国 外 电 子 与 通 信 教 材 系 列

PEARSON

英文版

射频微电子

（第二版）

RF Microelectronics, Second Edition



［美］ Behzad Razavi 著

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY http://www.phei.com.cn

射频微电子（第二版）（英文版）

RF Microelectronics, Second Edition

广受好评的射频微电子畅销书，针对最新的架构、电路和器件进行了全面扩展与更新。

无线通信和电力一样无处不在，但射频电路的设计仍然给工程师和研究者们带来巨大挑战。自本书第一版出版15年来，人们对高性能的不断追求使得射频技术爆炸式地增长。第二版与前版相比厚度翻倍，包含上百个例题与习题，更系统地讲述了射频电路与收发机分析与设计的基础知识和最新技术，针对低噪声放大器、混频器、振荡器和频率综合器等电路模块提出了若干新的设计方法与分析技术。最后新增一章，从WiFi的具体设计指标开始，为读者讲述了如何利用前面各章的知识，逐步完成晶体管级的射频收发机设计。

本书内容涵盖：

- 核心射频原理，包括噪声与非线性，并有机融合了模拟电路设计、微波理论与通信系统等内容。
- 从射频集成电路设计者的角度直观地描述了调制理论和无线通信标准。
- 各种收发机架构，如超外差式、滑动中频式、直接变频式、镜像抑制式和低中频式等。
- 低噪声放大器，包括Cascode共栅和共源等结构、噪声相消技术和电抗抵消技术等。
- 无源与有源混频器，包括增益与噪声的分析和新型混频器结构等。
- 压控振荡器，包括相位噪声机制、各种压控振荡器电路结构和关于噪声–功耗–调谐范围的设计权衡技术等。
- 无源元件，包括集成电感、MOS可变电容器和变压器等。
- 低相位噪声、低杂散锁相环的分析与设计。
- 整数N和分数N频率综合器，并介绍了分频器的设计。
- 功率放大器的原理与电路结构及发射机架构，包括极坐标调制式和反相调制式等。

Behzad Razavi

美国加州大学洛杉矶分校（UCLA）电机工程系教授，在长期的教学与科研工作中多次获得各种奖项。主要研究领域包括无线收发机、宽带数据通信电路和数据转换器等。IEEE会士，IEEE杰出教授，国际固态电路会议（ISSCC）50年以来排名前10位的作者之一，在模拟、射频和高速电路领域出版了7本专著。





策划编辑：马　岚
责任编辑：马　岚
责任美编：李　雯





ISBN 978-7-121-17603-6

定价：128.00 元

国外电子与通信教材系列

射频微电子

（第二版）

（英文版）

RF Microelectronics

Second Edition

［美］ Behzad Razavi 著

電子工業出版社
Publishing House of Electronics Industry
北京·BEIJING

内容简介

本书侧重系统级描述，综合了无线通信电路系统描述、器件特性及单元电路分析，讨论最新架构、电路和器件。第1章和第2章首先介绍射频电子学的基本概念和术语；第3章和第4章讨论通信系统层的建模、检测、多路存取等技术及无线标准；第5章讨论无线前端收发机的结构和集成电路的实现，第6章到第9章详细讨论了低噪声放大器和混频器、振荡器、频率综合器及功放器电路原理和分析方法。本书在改编过程中删除了英文原版中的少量习题。

本书既可作为高年级本科生或硕士研究生的双语教学教材，又可满足集成电路设计工作者进一步提高自身知识和设计技能的需求。

版权贸易合同登记号　图字：01-2012-4993

图书在版编目（CIP）数据
射频微电子 = RF Microelectronics：第2版：英文 /（美）拉扎维（Razavi, B.）著.
北京：电子工业出版社，2012.8
国外电子与通信教材系列
ISBN 978-7-121-17603-6

Ⅰ.①射…　Ⅱ.①拉…　Ⅲ.①射频－微电子技术－高等学校－教材－英文　Ⅳ.①TN4

中国版本图书馆 CIP 数据核字（2012）第158664号

策划编辑：马　岚
责任编辑：马　岚
印　　刷：
装　　订：三河市鑫金马印装有限公司
出版发行：电子工业出版社
　　　　　北京市海淀区万寿路173信箱　邮编：100036
开　　本：787 × 1092　1/16　　印张：58.5　　字数：1947千字
印　　次：2012年8月第1次印刷
定　　价：128.00元

凡所购买电子工业出版社的图书有缺损问题，请向购买书店调换；若书店售缺，请与本社发行部联系。联系及邮购电话：（010）88254888。
质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。
服务热线：（010）88258888。

导　读

*RF Microelectronics*一书的作者Behzad Razavi是美国加州大学洛杉矶分校终身教授，曾经在美国贝尔实验室和惠普实验室从事多年的射频电路设计工作，在射频电路领域有数十年的科研和教学经验。本书的第一版于1998年问世，经过不断的再版和翻译，成为射频电路设计领域的经典书籍。14年来，射频电路设计领域发生了巨大的变化，高集成度的无线设备和宽带的无线应用，促使科研人员在收发信机结构、电路形式及器件特性上，不断推陈出新。而且，新的电路分析方法及建模技术的成熟，使科研人员对射频电路的理解步入一个新的台阶。为反映这些变化，本书的第二版得以问世。

与旧版相比，新版在篇章结构与具体内容上都有显著变化，两者的内容重合度在10%左右。在新版著作中，作者通过大量的设计实例和问题讨论，帮助读者在学习射频电路整体分析方法的同时，了解射频电路设计中可能遇到的细节问题。同时，在新版著作中，作者也更加强调如何帮助读者掌握射频电路设计的基本方法，为此作者还特别增加了一章，用于指导读者如何一步一步地设计晶体管级的双频段WiFi收发机。

本书的具体内容可以概括如下。第2章介绍射频电路设计中的基本概念，其中增加了双端口网络S参数的定义和计算实例，为本书后续章节的分析打下基础。随后，第3章对无线通信的基本概念进行阐述，重点介绍数字调制方式及其相应的电路实现实例。第4章不仅介绍传统经典结构的各类收发信机，同时基于作者对射频电路最新发展趋势的跟踪，广受关注的新型收发信机结构也出现在新版著作中。值得一提的是，作者还通过问题讨论等方式，结合802.11a/g等具体无线通信标准，讲解了设计中需要注意的实际问题。本书的第5章至第12章，详尽介绍了无线收发信机中的各个子模块。与旧版相比，各子模块的分类方式有显著改进，作者也浓墨重彩地分析了各类新型模块技术，使读者能够及时地掌握射频电路设计的新趋势。新版还加入了无源器件的介绍与分析，使内容更趋完整。本书的第13章是收发信机设计实例，如前所述，本章内容是全书知识点的灵活运用，也是作者专注于设计方法传授的点睛之笔。

本书的内容体系基本涵盖了国内高校“通信基本电路”（亦称“高频电子线路”）专业基础课程的教学内容。但是，通过本人在上海交通大学电子工程系本科三年级的亲身教学实践（1学期64学时），发现本书与“通信基本电路”课程的教学大纲存在一定的不匹配之处。本书的内容相对于本科阶段的知识体系显得内容过于庞大，系统级的电路分析定性讲解有余，而单元电路的定量分析不足。因此，本书更适合作为理工类大专院校电子类专业研究生的课程教材。如果作为理工类大专院校通信、电子类本科生双语教学和全英文教学的教材，建议结合Thomas H. Lee的*Design of CMOS Radio-Frequency Integrated Circuits*（由电子工业出版社翻译出版），以便于学生掌握单元电路基础知识，为今后的科研打下扎实的基础。本书内容涵盖无线收发信机各个模块的介绍、分析和设计，并融入了Razavi教授数十年的电路设计经验，对从事射频电路设计的专业技术人员而言，更是一本不可多得的必备书籍。

甘小莺　副教授

上海交通大学电子工程系

PREFACE TO THE SECOND EDITION

In the 14 years since the first edition of this book, RF IC design has experienced a dramatic metamorphosis. Innovations in transceiver architectures, circuit topologies, and device structures have led to highly-integrated "radios" that span a broad spectrum of applications. Moreover, new analytical and modeling techniques have considerably improved our understanding of RF circuits and their underlying principles. A new edition was therefore due.

The second edition differs from the first in several respects:

1. I realized at the outset — three-and-a-half years ago — that simply adding "patches" to the first edition would not reflect today's RF microelectronics. I thus closed the first edition and began with a clean slate. The two editions have about 10% overlap.
2. I wanted the second edition to contain greater pedagogy, helping the reader understand both the fundamentals and the subtleties. I have thus incorporated hundreds of examples and problems.
3. I also wanted to teach design in addition to analysis. I have thus included step-by-step design procedures and examples. Furthermore, I have dedicated Chapter 13 to the step-by-step transistor-level design of a dual-band WiFi transceiver.
4. With the tremendous advances in RF design, some of the chapters have inevitably become longer and some have been split into two or more chapters. As a result, the second edition is nearly three times as long as the first.

Suggestions for Instructors and Students

The material in this book is much more than can be covered in one quarter or semester. The following is a possible sequence of the chapters that can be taught in one term with reasonable depth. Depending on the students' background and the instructor's preference, other combinations of topics can also be covered in one quarter or semester.

Chapter 1: Introduction to RF and Wireless Technology

This chapter provides the big picture and should be covered in about half an hour.

Chapter 2: Basic Concepts in RF Design

The following sections should be covered: General Considerations, Effects of Nonlinearity (the section on AM/PM Conversion can be skipped), Noise, and Sensitivity and Dynamic Range. (The

sections on Passive Impedance Transformation, Scattering Parameters, and Analysis of Nonlinear Dynamic Systems can be skipped.) This chapter takes about six hours of lecture.

Chapter 3: Communication Concepts

This chapter can be covered minimally in a quarter system — for example, Analog Modulation, Quadrature Modulation, GMSK Modulation, Multiple Access Techniques, and the IEEE802.11a/b/g Standard. In a semester system, the concept of signal constellations can be introduced and a few more modulation schemes and wireless standards can be taught. This chapter takes about two hours in a quarter system and three hours in a semester system.

Chapter 4: Transceiver Architectures

This chapter is relatively long and should be taught selectively. The following sections should be covered: General Considerations, Basic and Modern Heterodyne Receivers, Direct-Conversion Receivers, Image-Reject Receivers, and Direct-Conversion Transmitters. In a semester system, Low-IF Receivers and Heterodyne Transmitters can be covered as well. This chapter takes about eight hours in a quarter system and ten hours in a semester system.

Chapter 5: Low-Noise Amplifiers

The following sections should be covered: General Considerations, Problem of Input Matching, and LNA Topologies. A semester system can also include Gain Switching and Band Switching or High-IP_2 LNAs. This chapter takes about six hours in a quarter system and eight hours in a semester system.

Chapter 6: Mixers

The following sections should be covered: General Considerations, Passive Downconversion Mixers (the computation of noise and input impedance of voltage driven sampling mixers can be skipped), Active Downconversion Mixers, and Active Mixers with High IP_2. In a semester system, Active Mixers with Enhanced Transconductance, Active Mixers with Low Flicker Noise, and Upconversion Mixers can also be covered. This chapter takes about eight hours in a quarter system and ten hours in a semester system.

Chapter 7: Passive Devices

This chapter may not fit in a quarter system. In a semester system, about three hours can be spent on basic inductor structures and loss mechanisms and MOS varactors.

Chapter 8: Oscillators

This is a long chapter and should be taught selectively. The following sections should be covered: Basic Principles, Cross-Coupled Oscillator, Voltage-Controlled Oscillators, Low-Noise VCOs. In a quarter system, there is little time to cover phase noise. In a semester system, both approaches to phase

noise analysis can be taught. This chapter takes about six hours in a quarter system and eight hours in a semester system.

Chapter 9: Phase–Locked Loops

This chapter forms the foundation for synthesizers. In fact, if taught carefully, this chapter naturally teaches integer-N synthesizers, allowing a quarter system to skip the next chapter. The following sections should be covered: Basic Concepts, Type-I PLLs, Type-II PLLs, and PFD/CP Nonidealities. A semester system can also include Phase Noise in PLLs and Design Procedure. This chapter takes about four hours in a quarter system and six hours in a semester system.

Chapter 10: Integer–N Synthesizers

This chapter is likely sacrificed in a quarter system. A semester system can spend about four hours on Spur Reduction Techniques and Divider Design.

Chapter 11: Fractional–N Synthesizers

This chapter is likely sacrificed in a quarter system. A semester system can spend about four hours on Randomization and Noise Shaping. The remaining sections may be skipped.

Chapter 12: Power Amplifiers

This is a long chapter and, unfortunately, is often sacrificed for other chapters. If coverage is desired, the following sections may be taught: General Considerations, Classification of Power Amplifiers, High-Efficiency Power Amplifiers, Cascode Output Stages, and Basic Linearization Techniques. These topics take about four hours of lecture. Another four hours can be spent on Doherty Power Amplifier, Polar Modulation, and Outphasing.

Chapter 13: Transceiver Design Example

This chapter provides a step-by-step design of a dual-band transceiver. It is possible to skip the state-of-the-art examples in Chapters 5, 6, and 8 to allow some time for this chapter. The system-level derivations may still need to be skipped. The RX, TX, and synthesizer transistor-level designs can be covered in about four hours.

A solutions manual is available for instructors via the Pearson Higher Education Instructor Resource Center web site: pearsonhighered.com/irc; and a set of Powerpoint slides is available for instructors at informit.com/razavi. Additional problems will be posted on the book's website (informit.com/razavi).①

— Behzad Razavi
July 2011

---

① 登录华信教育资源网（http://www.hxedu.com.cn）可下载本书相关资源。采用本书作为教材的教师可获得本书配套教辅和习题解答（英文版）。详见书末所附“教学支持说明”。

PREFACE TO THE FIRST EDITION

The annual worldwide sales of cellular phones has exceeded $2.5B. With 4.5 million customers, home satellite networks comprise a $2.5B industry. The global positioning system is expected to become a $5B market by the year 2000. In Europe, the sales of equipment and services for mobile communications will reach $30B by 1998. The statistics are overwhelming.

The radio frequency (RF) and wireless market has suddenly expanded to unimaginable dimensions. Devices such as pagers, cellular and cordless phones, cable modems, and RF identification tags are rapidly penetrating all aspects of our lives, evolving from luxury items to indispensable tools. Semiconductor and system companies, small and large, analog and digital, have seen the statistics and are striving to capture their own market share by introducing various RF products.

RF design is unique in that it draws upon many disciplines unrelated to integrated circuits (ICs). The RF knowledge base has grown for almost a century, creating a seemingly endless body of literature for the novice.

This book deals with the analysis and design of RF integrated circuits and systems. Providing a systematic treatment of RF electronics in a tutorial language, the book begins with the necessary background knowledge from microwave and communication theory and leads the reader to the design of RF transceivers and circuits. The text emphasizes both architecture and circuit level issues with respect to monolithic implementation in VLSI technologies. The primary focus is on bipolar and CMOS design, but most of the concepts can be applied to other technologies as well. The reader is assumed to have a basic understanding of analog IC design and the theory of signals and systems.

The book consists of nine chapters. Chapter 1 gives a general introduction, posing questions and providing motivation for subsequent chapters. Chapter 2 describes basic concepts in RF and microwave design, emphasizing the effects of nonlinearity and noise.

Chapters 3 and 4 take the reader to the communication system level, giving an overview of modulation, detection, multiple access techniques, and wireless standards. While initially appearing to be unnecessary, this material is in fact essential to the concurrent design of RF circuits and systems.

Chapter 5 deals with transceiver architectures, presenting various receiver and transmitter topologies along with their merits and drawbacks. This chapter also includes a number of case studies that exemplify the approaches taken in actual RF products.

Chapters 6 through 9 address the design of RF building blocks: low-noise amplifiers and mixers, oscillators, frequency synthesizers, and power amplifiers, with particular attention to minimizing the number of off-chip components. An important goal of these chapters is to demonstrate how the system requirements define the parameters of the circuits and how the performance of each circuit impacts that of the overall transceiver.

I have taught approximately 80% of the material in this book in a 4-unit graduate course at UCLA. Chapters 3, 4, 8, and 9 had to be shortened in a ten-week quarter, but in a semester system they can be covered more thoroughly.

Much of my RF design knowledge comes from interactions with colleagues. Helen Kim, Ting-Ping Liu, and Dan Avidor of Bell Laboratories, and David Su and Andrew Gzegorek of Hewlett-Packard Laboratories have contributed to the material in this book in many ways. The text was also reviewed by a number of experts: Stefan Heinen (Siemens), Bart Jansen (Hewlett-Packard), Ting-Ping Liu (Bell Labs), John Long (University of Toronto), Tadao Nakagawa (NTT), Gitty Nasserbakht (Texas Instruments), Ted Rappaport (Virginia Tech), Tirdad Sowlati (Gennum), Trudy Stetzler (Bell Labs), David Su (Hewlett-Packard), and Rick Wesel (UCLA). In addition, a number of UCLA students, including Farbod Behbahani, Hooman Darabi, John Leete, and Jacob Rael, "test drove" various chapters and provided useful feedback. I am indebted to all of the above for their kind assistance.

I would also like to thank the staff at Prentice Hall, particularly Russ Hall, Maureen Diana, and Kerry Riordan for their support.

— Behzad Razavi
July 1997

ACKNOWLEDGMENTS

I have been fortunate to benefit from the support of numerous people during the writing, review, and production phases of this book. I would like to express my thanks here.

Even after several rounds of self-editing, it is possible that typos or subtle mistakes have eluded the author. Sometimes, an explanation that is clear to the author may not be so to the reader. And, occasionally, the author may have missed a point or a recent development. A detailed review of the book by others thus becomes necessary. The following individuals meticulously reviewed various chapters, discovered my mistakes, and made valuable suggestions:

Ali Afsahi (Broadcom)
Pietro Andreani (Lund University)
Ashkan Borna (UC Berkeley)
Jonathan Borremans (IMEC)
Debopriyo Chowdhury (UC Berkeley)
Matteo Conta (Consultant)
Ali Homayoun (UCLA)
Velntina del Lattorre (Consultant)
Jane Gu (University of Florida)
Peng Han (Beken)
Pavan Hanumolu (Oregon State University)
Daquan Huang (Texas Instruments)
Sy-Chyuan Hwu (UCLA)
Amin Jahanian (UCI)
Jithin Janardhan (UCLA)
Shinwon Kang (UC Berkeley)
Iman Khajenasiri (Sharif University of Technology)
Yanghyo Kim (UCLA)
Abbas Komijani (Atheros)
Tai-Cheng Lee (National Taiwan University)
Antonio Liscidini (University of Pavia)
Shen-Iuan Liu (National Taiwan University)
Xiaodong Liu (Lund University)
Jian Hua Lu (UCLA)
Howard Luong (Hong Kong University of Science and Technology)
Elvis Mak (University of Macau)
Rabih Makarem (Atheros)
Rui Martins (University of Macau)
Andrea Mazzanti (University of Pavia)
Karthik Natarajan (University of Washington)
Nitin Nidhi (UCLA)
Joung Park (UCLA)
Paul Park (Atheros)
Stefano Pellerano (Intel)
Jafar Savoj (Xilinx)
Parmoon Seddighrad (University of Washington)
Alireza Shirvani (Ralink)
Tirdad Sowlati (Qualcomm)
Francesco Svelto (University of Pavia)
Enrico Temporiti (ST Microelectronics)
Federico Vecchi (University of Pavia)
Vijay Viswam (Lund University)
Vidojkovic Vojkan (IMEC)
Ning Wang (UCLA)
Weifeng Wang (Beken)
Zhi Gong Wang (Southeast University)
Marco Zanuso (UCLA)
Yunfeng Zhao (Beken)
Alireza Zolfaghari (Broadcom)

I am thankful for their enthusiastic, organized, and to-the-point reviews.

The book's production was proficiently managed by the staff at Prentice Hall, including Bernard Goodwin and Julie Nahil. I would like to thank both.

As with my other books, my wife, Angelina, typed the entire second edition in Latex and selflessly helped me in this three-and-a-half-year endeavor. I am grateful to her.

— *Behzad Razavi*

ABOUT THE AUTHOR

Behzad Razavi received the BSEE degree from Sharif University of Technology in 1985 and MSEE and PhDEE degrees from Stanford University in 1988 and 1992, respectively. He was with AT&T Bell Laboratories and Hewlett-Packard Laboratories until 1996. Since 1996, he has been associate professor and, subsequently, professor of electrical engineering at University of California, Los Angeles. His current research includes wireless transceivers, frequency synthesizers, phase-locking and clock recovery for high-speed data communications, and data converters.

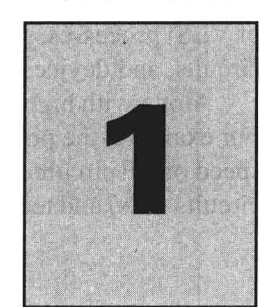
Professor Razavi was an adjunct professor at Princeton University from 1992 to 1994, and at Stanford University in 1995. He served on the Technical Program Committees of the International Solid-State Circuits Conference (ISSCC) from 1993 to 2002 and VLSI Circuits Symposium from 1998 to 2002. He has also served as guest editor and associate editor of the IEEE Journal of Solid-State Circuits, IEEE Transactions on Circuits and Systems, and International Journal of High Speed Electronics.

Professor Razavi received the Beatrice Winner Award for Editorial Excellence at the 1994 ISSCC; the best paper award at the 1994 European Solid-State Circuits Conference; the best panel award at the 1995 and 1997 ISSCC; the TRW Innovative Teaching Award in 1997; the best paper award at the IEEE Custom Integrated Circuits Conference (CICC) in 1998; and McGraw-Hill First Edition of the Year Award in 2001. He was the co-recipient of both the Jack Kilby Outstanding Student Paper Award and the Beatrice Winner Award for Editorial Excellence at the 2001 ISSCC. He received the Lockheed Martin Excellence in Teaching Award in 2006; the UCLA Faculty Senate Teaching Award in 2007; and the CICC Best Invited Paper Award in 2009. He was also recognized as one of the top ten authors in the fifty-year history of ISSCC. He received the IEEE Donald Pederson Award in Solid-State Circuits in 2012.

Professor Razavi is an IEEE Distinguished Lecturer, a Fellow of IEEE, and the author of Principles of Data Conversion System Design, RF Microelectronics, First Edition (translated to Chinese, Japanese, and Korean), Design of Analog CMOS Integrated Circuits (translated to Chinese, Japanese, and Korean), Design of Integrated Circuits for Optical Communications, and Fundamentals of Microelectronics (translated to Korean and Portuguese), and the editor of Monolithic Phase-Locked Loops and Clock Recovery Circuits and Phase-Locking in High-Performance Systems.

CONTENTS

CHAPTER



INTRODUCTION TO RF AND WIRELESS TECHNOLOGY

Compare two RF transceivers designed for cell phones:

> *"A 2.7-V GSM RF Transceiver IC" [1] (published in 1997)*
>
> *"A Single-Chip 10-Band WCDMA/HSDPA 4-Band GSM/EDGE SAW-Less CMOS Receiver with DigRF 3G Interface and +90-dBm* IIP_2*" [2] (published in 2009)*

Why is the latter much more complex than the former? Does the latter have a higher performance or only greater functionality? Which one costs more? Which one consumes a higher power? What do all the acronyms GSM, WCDMA, HSDPA, EDGE, SAW, and IIP_2 mean? Why do we care?

The field of RF communication has grown rapidly over the past two decades, reaching far into our lives and livelihood. Our cell phones serve as an encyclopedia, a shopping terminus, a GPS guide, a weather monitor, and a telephone—all thanks to their wireless communication devices. We can now measure a patient's brain or heart activity and transmit the results wirelessly, allowing the patient to move around untethered. We use RF devices to track merchandise, pets, cattle, children, and convicts.

1.1 A WIRELESS WORLD

Wireless communication has become almost as ubiquitous as electricity; our refrigerators and ovens may not have a wireless device at this time, but it is envisioned that our homes will eventually incorporate a wireless network that controls every device and appliance. High-speed wireless links will allow seamless connections among our laptops, digital cameras, camcorders, cell phones, printers, TVs, microwave ovens, etc. Today's WiFi and Bluetooth connections are simple examples of such links.

How did wireless communication take over the world? A confluence of factors has contributed to this explosive growth. The principal reason for the popularity of wireless

communication is the ever-decreasing cost of electronics. Today's cell phones cost about the same as those a decade ago but they offer many more functions and features: many frequency bands and communication modes, WiFi, Bluetooth, GPS, computing, storage, a digital camera, and a user-friendly interface. This affordability finds its roots in *integration*, i.e., how much functionality can be placed on a single chip—or, rather, how few components are left off-chip. The integration, in turn, owes its steady rise to (1) the scaling of VLSI processes, particularly, CMOS technology, and (2) innovations in RF architectures, circuits, and devices.

Along with higher integration levels, the performance of RF circuits has also improved. For example, the power consumption necessary for a given function has decreased and the speed of RF circuits has increased. Figure 1.1 illustrates some of the trends in RF integrated circuits (ICs) and technology for the past two decades. The minimum feature size of CMOS

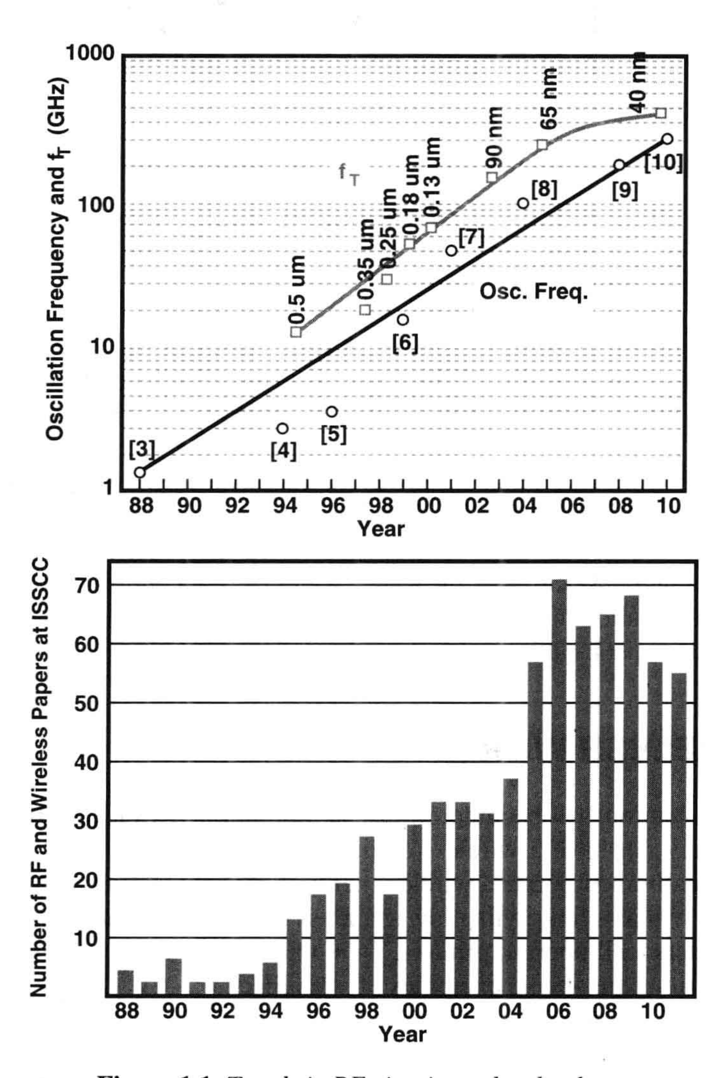


Figure 1.1 *Trends in RF circuits and technology.*

technology has fallen from 0.5 μm to 40 nm, the transit frequency,[1] f_T, of NMOS devices has risen from about 12 GHz to several hundred gigahertz, and the speed of RF oscillators has gone from 1.2 GHz to 300 GHz. Also shown is the number of RF and wireless design papers presented at the International Solid-State Circuits Conference (ISSCC) each year, revealing the fast-growing activity in this field.

1.2 RF DESIGN IS CHALLENGING

Despite many decades of work on RF and microwave theory and two decades of research on RF ICs, the design and implementation of RF circuits and transceivers remain challenging. This is for three reasons. First, as shown in Fig. 1.2, RF design draws upon a multitude of disciplines, requiring a good understanding of fields that are seemingly irrelevant to integrated circuits. Most of these fields have been under study for more than half a century, presenting a massive body of knowledge to a person entering RF IC design. One objective of this book is to provide the necessary background from these disciplines without overwhelming the reader.

Second, RF circuits and transceivers must deal with numerous trade-offs, summarized in the "RF design hexagon" of Fig. 1.3. For example, to lower the noise of a front-end amplifier, we must consume a greater power or sacrifice linearity. We will encounter these trade-offs throughout this book.

Third, the demand for higher performance, lower cost, and greater functionality continues to present new challenges. The early RF IC design work in the 1990s strove to integrate *one* transceiver—perhaps along with the digital baseband processor—on a single chip. Today's efforts, on the other hand, aim to accommodate multiple transceivers operating in different frequency bands for different wireless standards (e.g., Bluetooth, WiFi, GPS, etc.). The two papers mentioned at the beginning of this chapter exemplify this trend. It is interesting to note that the silicon chip area of early single-transceiver systems was

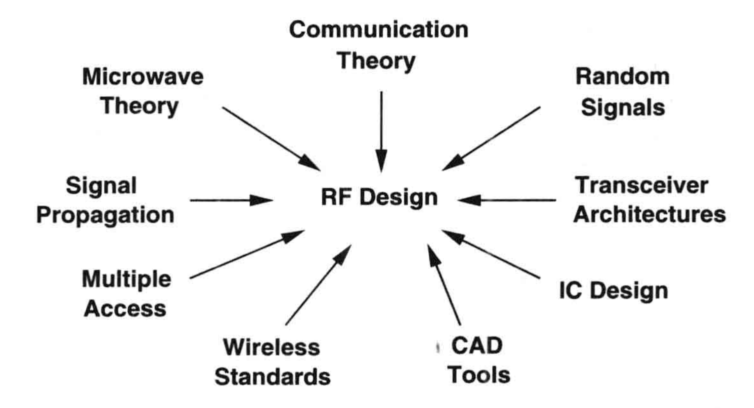


Figure 1.2 *Various disciplines necessary in RF design.*

1. The transit frequency is defined as the frequency at which the small-signal current gain of a device falls to unity.

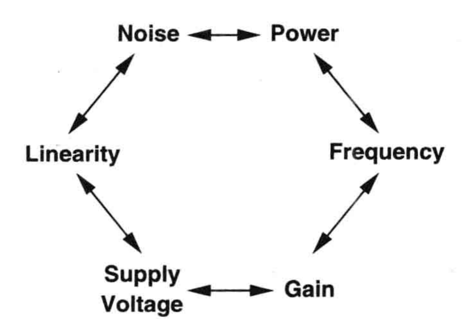


Figure 1.3 *RF design hexagon.*

dominated by the digital baseband processor, allowing RF and analog designers some latitude in the choice of their circuit and device topologies. In today's designs, however, the multiple transceivers tend to occupy a *larger* area than the baseband processor, requiring that RF and analog sections be designed with much care about their area consumption. For example, while on-chip spiral inductors (which have a large footprint) were utilized in abundance in older systems, they are now used only sparingly.

1.3 THE BIG PICTURE

The objective of an RF transceiver is to transmit and receive information. We envision that the transmitter (TX) somehow processes the voice or data signal and applies the result to the antenna [Fig. 1.4(a)]. Similarly, the receiver (RX) senses the signal picked up by the antenna and processes it so as to reconstruct the original voice or data information. Each black box in Fig. 1.4(a) contains a great many functions, but we can readily make two observations: (1) the TX must drive the antenna with a high power level so that the transmitted signal is strong enough to reach far distances, and (2) the RX may sense a small signal (e.g., when a cell phone is used in the basement of a building) and must first amplify the signal with low noise. We now architect our transceiver as shown in Fig. 1.4(b), where the signal to be transmitted is first applied to a "modulator" or "upconverter" so that its center frequency goes from zero to, say, $f_c = 2.4$ GHz. The result drives the antenna through a "power amplifier" (PA). On the receiver side, the signal is sensed by a "low-noise amplifier" (LNA) and subsequently by a "downconverter" or "demodulator" (also known as a "detector").

The upconversion and downconversion paths in Fig. 1.4(b) are driven by an oscillator, which itself is controlled by a "frequency synthesizer." Figure 1.4(c) shows the overall transceiver.[2] The system looks deceptively simple, but we will need the next 900 pages to cover its RF sections. And perhaps another 900 pages to cover the analog-to-digital and digital-to-analog converters.

2. In some cases, the modulator and the upconverter are one and the same. In some other cases, the modulation is performed in the digital domain before upconversion. Most receivers demodulate and detect the signal digitally, requiring only a downconverter in the analog domain.

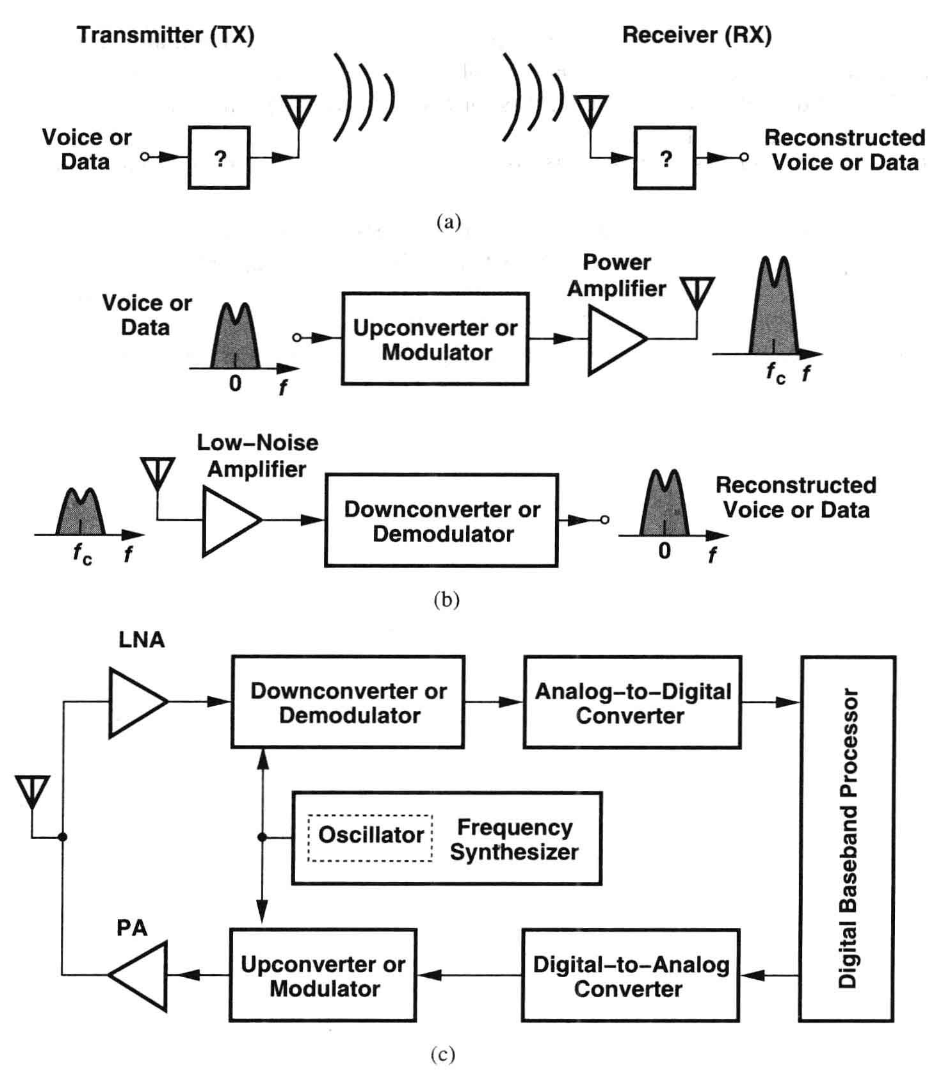


Figure 1.4 *(a) Simple view of RF communication, (b) more complete view, (c) generic RF transceiver.*

REFERENCES

[1] T. Yamawaki et al., "A 2.7-V GSM RF Transceiver IC," *IEEE J. Solid-State Circuits,* vol. 32, pp. 2089–2096, Dec. 1997.

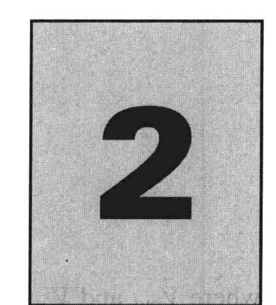
[2] D. Kaczman et al., "A Single-Chip 10-Band WCDMA/HSDPA 4-Band GSM/EDGE SAW-less CMOS Receiver with DigRF 3G Interface and +90-dBm IIP2," *IEEE J. Solid-State Circuits,* vol. 44, pp. 718–739, March 2009.

[3] M. Banu, "MOS Oscillators with Multi-Decade Tuning Range and Gigahertz Maximum Speed," *IEEE J. Solid-State Circuits,* vol. 23, pp. 474–479, April 1988.

[4] B. Razavi et al., "A 3-GHz 25-mW CMOS Phase-Locked Loop," *Dig. of Symposium on VLSI Circuits*, pp. 131–132, June 1994.

[5] M. Soyuer et al., "A 3-V 4-GHz nMOS Voltage-Controlled Oscillator with Integrated Resonator," *IEEE J. Solid-State Circuits,* vol. 31, pp. 2042–2045, Dec. 1996.
[6] B. Kleveland et al., "Monolithic CMOS Distributed Amplifier and Oscillator," *ISSCC Dig. Tech. Papers,* pp. 70–71, Feb. 1999.
[7] H. Wang, "A 50-GHz VCO in 0.25-μm CMOS," *ISSCC Dig. Tech. Papers,* pp. 372–373, Feb. 2001.
[8] L. Franca-Neto, R. Bishop, and B. Bloechel, "64 GHz and 100 GHz VCOs in 90 nm CMOS Using Optimum Pumping Method," ISSCC Dig. Tech. Papers, pp. 444–445, Feb. 2004.
[9] E. Seok et al., "A 410GHz CMOS Push-Push Oscillator with an On-Chip Patch Antenna" *ISSCC Dig. Tech. Papers,* pp. 472–473, Feb. 2008.
[10] B. Razavi, "A 300-GHz Fundamental Oscillator in 65-nm CMOS Technology," *Symposium on VLSI Circuits Dig. Of Tech. Papers,* pp. 113–114, June 2010.

CHAPTER



BASIC CONCEPTS IN RF DESIGN

RF design draws upon many concepts from a variety of fields, including signals and systems, electromagnetics and microwave theory, and communications. Nonetheless, RF design has developed its own analytical methods and its own language. For example, while the nonlinear behavior of analog circuits may be characterized by "harmonic distortion," that of RF circuits is quantified by very different measures.

This chapter deals with general concepts that prove essential to the analysis and design of RF circuits, closing the gaps with respect to other fields such as analog design, microwave theory, and communication systems. The outline is shown below.

Nonlinearity	Noise	Impedance Transformation
■ Harmonic Distortion	■ Noise Spectrum	■ Series–Parallel Conversion
■ Compression	■ Device Noise	■ Matching Networks
■ Intermodulation	■ Noise in Circuits	■ S-Parameters
■ Dynamic Nonlinear Systems		

2.1 GENERAL CONSIDERATIONS

2.1.1 Units in RF Design

RF design has traditionally employed certain units to express gains and signal levels. It is helpful to review these units at the outset so that we can comfortably use them in our subsequent studies.

The voltage gain, V_{out}/V_{in}, and power gain, P_{out}/P_{in}, are expressed in decibels (dB):

$$A_V|_{\text{dB}} = 20 \log \frac{V_{out}}{V_{in}} \tag{2.1}$$

$$A_P|_{\text{dB}} = 10 \log \frac{P_{out}}{P_{in}}. \tag{2.2}$$

These two quantities are equal (in dB) only if the input and output voltages appear across *equal* impedances. For example, an amplifier having an input resistance of R_0 (e.g., 50 Ω) and driving a load resistance of R_0 satisfies the following equation:

$$A_P|_{\text{dB}} = 10 \log \frac{\dfrac{V_{out}^2}{R_0}}{\dfrac{V_{in}^2}{R_0}} \tag{2.3}$$

$$= 20 \log \frac{V_{out}}{V_{in}} \tag{2.4}$$

$$= A_V|_{\text{dB}}, \tag{2.5}$$

where V_{out} and V_{in} are rms values. In many RF systems, however, this relationship does not hold because the input and output impedances are not equal.

The absolute signal levels are often expressed in dBm rather than in watts or volts. Used for power quantities, the unit dBm refers to "dB's above 1 mW." To express the signal power, P_{sig}, in dBm, we write

$$P_{sig}|_{\text{dBm}} = 10 \log \left(\frac{P_{sig}}{1\text{ mW}} \right). \tag{2.6}$$

Example 2.1

An amplifier senses a sinusoidal signal and delivers a power of 0 dBm to a load resistance of 50 Ω. Determine the peak-to-peak voltage swing across the load.

Solution:

Since 0 dBm is equivalent to 1 mW, for a sinusoidal having a peak-to-peak amplitude of V_{pp} and hence an rms value of $V_{pp}/(2\sqrt{2})$, we write

$$\frac{V_{pp}^2}{8R_L} = 1\text{ mW}, \tag{2.7}$$

where $R_L = 50\ \Omega$. Thus,

$$V_{pp} = 632\text{ mV}. \tag{2.8}$$

This is an extremely useful result, as demonstrated in the next example.

Example 2.2

A GSM receiver senses a narrowband (modulated) signal having a level of −100 dBm. If the front-end amplifier provides a voltage gain of 15 dB, calculate the peak-to-peak voltage swing at the output of the amplifier.

Example 2.2 (Continued)

Solution:

Since the amplifier output *voltage* swing is of interest, we first convert the received signal level to voltage. From the previous example, we note that $-100\,\text{dBm}$ is $100\,\text{dB}$ below $632\,\text{mV}_{\text{pp}}$. Also, 100 dB for voltage quantities is equivalent to 10^5. Thus, $-100\,\text{dBm}$ is equivalent to $6.32\,\mu\text{V}_{\text{pp}}$. This input level is amplified by 15 dB (≈ 5.62), resulting in an output swing of $35.5\,\mu\text{V}_{\text{pp}}$.

The reader may wonder why the output *voltage* of the amplifier is of interest in the above example. This may occur if the circuit following the amplifier does not present a 50-Ω input impedance, and hence the power gain and voltage gain are not equal in dB. In fact, the next stage may exhibit a purely *capacitive* input impedance, thereby requiring no signal "power." This situation is more familiar in analog circuits wherein one stage drives the gate of the transistor in the next stage. As explained in Chapter 5, in most integrated RF systems, we prefer voltage quantities to power quantities so as to avoid confusion if the input and output impedances of cascade stages are unequal or contain negligible real parts.

The reader may also wonder why we were able to assume 0 dBm is equivalent to $632\,\text{mV}_{\text{pp}}$ in the above example even though the signal is not a pure sinusoid. After all, only for a sinusoid can we assume that the rms value is equal to the peak-to-peak value divided by $2\sqrt{2}$. Fortunately, for a narrowband 0-dBm signal, it is still possible to approximate the (average) peak-to-peak swing as 632 mV.

Although dBm is a unit of power, we sometimes use it at interfaces that do not necessarily entail power transfer. For example, consider the case shown in Fig. 2.1(a), where the LNA drives a purely-capacitive load with a $632\text{-mV}_{\text{pp}}$ swing, delivering no average power. We mentally attach an ideal voltage buffer to node X and drive a 50-Ω load [Fig. 2.1(b)]. We then say that the signal at node X has a level of 0 dBm, tacitly meaning that *if* this signal were applied to a 50-Ω load, *then* it would deliver 1 mW.

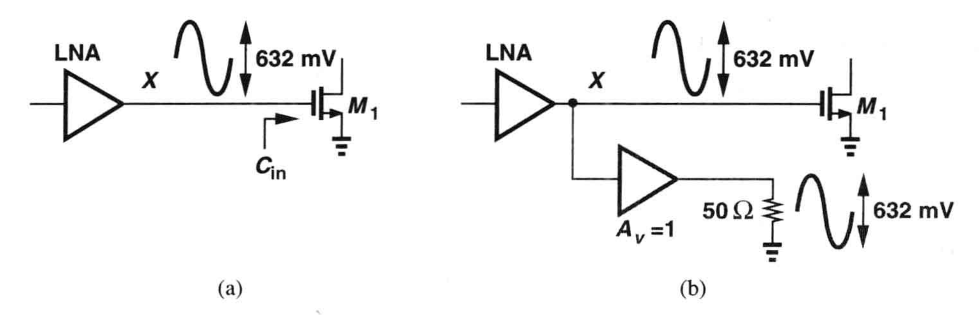


Figure 2.1 *(a) LNA driving a capacitive impedance, (b) use of fictitious buffer to visualize the signal level in dBm.*

2.1.2 Time Variance

A system is linear if its output can be expressed as a linear combination (superposition) of responses to individual inputs. More specifically, if the outputs in response to inputs $x_1(t)$

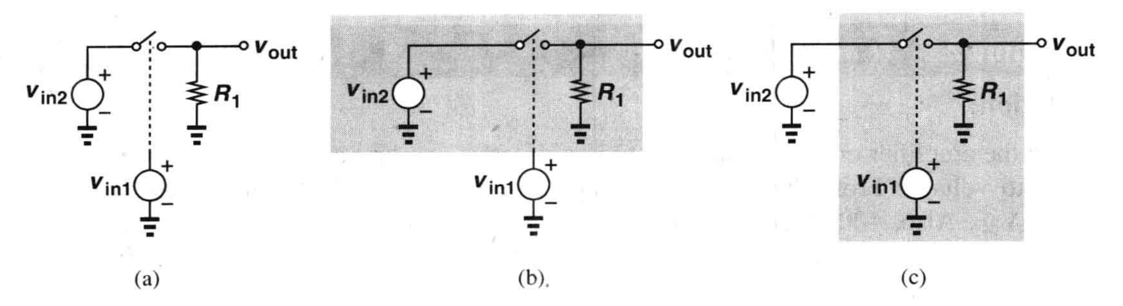


Figure 2.2 *(a) Simple switching circuit, (b) system with V_{in1} as the input, (c) system with V_{in2} as the input.*

and $x_2(t)$ can be respectively expressed as

$$y_1(t) = f[x_1(t)] \tag{2.9}$$

$$y_2(t) = f[x_2(t)], \tag{2.10}$$

then,

$$ay_1(t) + by_2(t) = f[ax_1(t) + bx_2(t)], \tag{2.11}$$

for arbitrary values of a and b. Any system that does not satisfy this condition is nonlinear. Note that, according to this definition, nonzero initial conditions or dc offsets also make a system nonlinear, but we often relax the rule to accommodate these two effects.

Another attribute of systems that may be confused with nonlinearity is time variance. A system is time-invariant if a time shift in its input results in the same time shift in its output. That is, if $y(t) = f[x(t)]$, then $y(t - \tau) = f[x(t - \tau)]$ for arbitrary τ.

As an example of an RF circuit in which time variance plays a critical role and must not be confused with nonlinearity, let us consider the simple switching circuit shown in Fig. 2.2(a). The control terminal of the switch is driven by $v_{in1}(t) = A_1 \cos\omega_1 t$ and the input terminal by $v_{in2}(t) = A_2 \cos\omega_2 t$. We assume the switch is on if $v_{in1} > 0$ and off otherwise. Is this system nonlinear or time-variant? If, as depicted in Fig. 2.2(b), the input of interest is v_{in1} (while v_{in2} is part of the system and still equal to $A_2 \cos\omega_2 t$), then the system is nonlinear because the control is only sensitive to the polarity of v_{in1} and independent of its amplitude. This system is also time-variant because the output depends on v_{in2}. For example, if v_{in1} is constant and positive, then $v_{out}(t) = v_{in2}(t)$, and if v_{in1} is constant and negative, then $v_{out}(t) = 0$ (why?).

Now consider the case shown in Fig. 2.2(c), where the input of interest is v_{in2} (while v_{in1} remains part of the system and still equal to $A_1 \cos\omega_1 t$). This system is linear with respect to v_{in2}. For example, doubling the amplitude of v_{in2} directly doubles that of v_{out}. The system is also time-variant due to the effect of v_{in1}.

Example 2.3

Plot the output waveform of the circuit in Fig. 2.2(a) if $v_{in1} = A_1 \cos\omega_1 t$ and $v_{in2} = A_2 \cos(1.25\omega_1 t)$.

Example 2.3 *(Continued)*

Solution:

As shown in Fig. 2.3, v_{out} tracks v_{in2} if $v_{in1} > 0$ and is pulled down to zero by R_1 if $v_{in1} < 0$. That is, v_{out} is equal to the product of v_{in2} and a square wave toggling between 0 and 1.

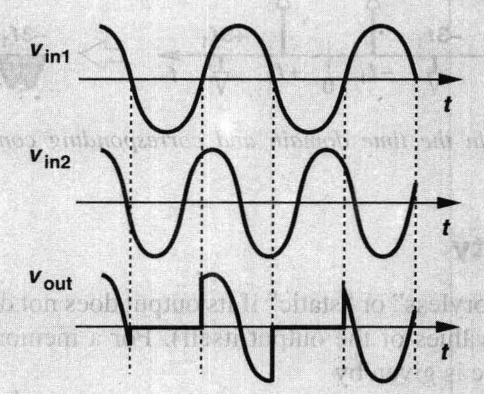


Figure 2.3 *Input and output waveforms.*

The circuit of Fig. 2.2(a) is an example of RF "mixers." We will study such circuits in Chapter 6 extensively, but it is important to draw several conclusions from the above study. First, statements such as "switches are nonlinear" are ambiguous. Second, a linear system *can* generate frequency components that do not exist in the input signal—the system only need be time-variant. From Example 2.3,

$$v_{out}(t) = v_{in2}(t) \cdot S(t), \tag{2.12}$$

where $S(t)$ denotes a square wave toggling between 0 and 1 with a frequency of $f_1 = \omega_1/(2\pi)$. The output spectrum is therefore given by the convolution of the spectra of $v_{in2}(t)$ and $S(t)$. Since the spectrum of a square wave is equal to a train of impulses whose amplitudes follow a sinc envelope, we have

$$V_{out}(f) = V_{in2}(f) * \sum_{n=-\infty}^{+\infty} \frac{\sin(n\pi/2)}{n\pi} \delta\left(f - \frac{n}{T_1}\right) \tag{2.13}$$

$$= \sum_{n=-\infty}^{+\infty} \frac{\sin(n\pi/2)}{n\pi} V_{in2}\left(f - \frac{n}{T_1}\right), \tag{2.14}$$

where $T_1 = 2\pi/\omega_1$. This operation is illustrated in Fig. 2.4 for a V_{in2} spectrum located around zero frequency.[1]

1. It is helpful to remember that, for $n = 1$, each impulse in the above summation has an area of $1/\pi$ and the corresponding sinusoid, a *peak amplitude* of $2/\pi$.

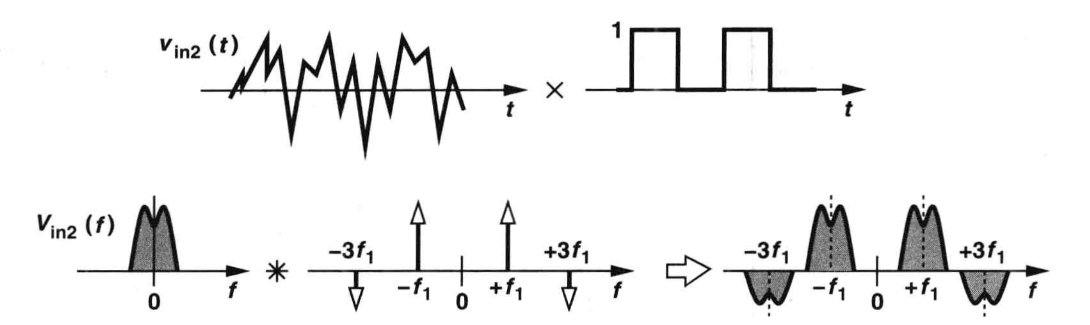


Figure 2.4 *Multiplication in the time domain and corresponding convolution in the frequency domain.*

2.1.3 Nonlinearity

A system is called "memoryless" or "static" if its output does not depend on the past values of its input (or the past values of the output itself). For a memoryless linear system, the input/output characteristic is given by

$$y(t) = \alpha x(t), \tag{2.15}$$

where α is a function of time if the system is time-variant [e.g., Fig. 2.2(c)]. For a memoryless nonlinear system, the input/output characteristic can be approximated with a polynomial,

$$y(t) = \alpha_0 + \alpha_1 x(t) + \alpha_2 x^2(t) + \alpha_3 x^3(t) + \cdots, \tag{2.16}$$

where α_j may be functions of time if the system is time-variant. Figure 2.5 shows a common-source stage as an example of a memoryless nonlinear circuit (at low frequencies). If M_1 operates in the saturation region and can be approximated as a square-law device, then

$$V_{out} = V_{DD} - I_D R_D \tag{2.17}$$

$$= V_{DD} - \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{in} - V_{TH})^2 R_D. \tag{2.18}$$

In this idealized case, the circuit displays only second-order nonlinearity.

The system described by Eq. (2.16) has "odd symmetry" if $y(t)$ is an odd function of $x(t)$, i.e., if the response to $-x(t)$ is the negative of that to $+x(t)$. This occurs if $\alpha_j = 0$ for even j. Such a system is sometimes called "balanced," as exemplified by the differential

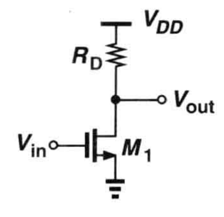


Figure 2.5 *Common-source stage.*

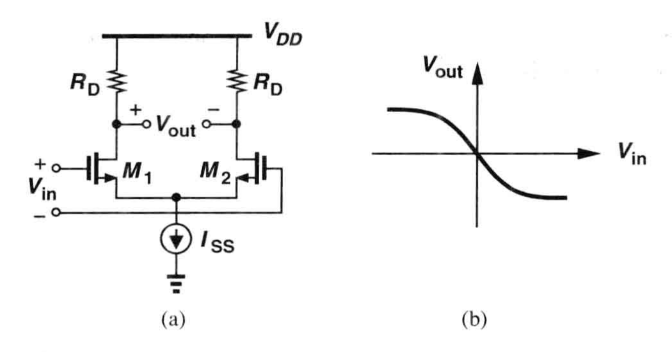


Figure 2.6 *(a) Differential pair and (b) its input/output characteristic.*

pair shown in Fig. 2.6(a). Recall from basic analog design that by virtue of symmetry, the circuit exhibits the characteristic depicted in Fig. 2.6(b) if the differential input varies from very negative values to very positive values.

Example 2.4

For square-law MOS transistors operating in saturation, the characteristic of Fig. 2.6(b) can be expressed as [1]

$$V_{out} = -\frac{1}{2}\mu_n C_{ox}\frac{W}{L}V_{in}\sqrt{\frac{4I_{SS}}{\mu_n C_{ox}\frac{W}{L}} - V_{in}^2}R_D. \tag{2.19}$$

If the differential input is small, approximate the characteristic by a polynomial.

Solution:

Factoring $4I_{SS}/(\mu_n C_{ox}W/L)$ out of the square root and assuming

$$V_{in}^2 \ll \frac{4I_{SS}}{\mu_n C_{ox}\frac{W}{L}}, \tag{2.20}$$

we use the approximation $\sqrt{1-\epsilon} \approx 1 - \epsilon/2$ to write

$$V_{out} \approx -\sqrt{\mu_n C_{ox}\frac{W}{L}I_{SS}}V_{in}\left(1 - \frac{\mu_n C_{ox}\frac{W}{L}}{8I_{SS}}V_{in}^2\right)R_D \tag{2.21}$$

$$\approx -\sqrt{\mu_n C_{ox}\frac{W}{L}I_{SS}}R_D V_{in} + \frac{\left(\mu_n C_{ox}\frac{W}{L}\right)^{3/2}}{8\sqrt{I_{SS}}}R_D V_{in}^3. \tag{2.22}$$

The first term on the right-hand side represents linear operation, revealing the small-signal voltage gain of the circuit ($-g_m R_D$). Due to symmetry, even-order nonlinear terms are absent. Interestingly, square-law devices yield a *third-order* characteristic in this case. We return to this point in Chapter 5.

A system is called "dynamic" if its output depends on the past values of its input(s) or output(s). For a linear, time-invariant, dynamic system,

$$y(t) = h(t) * x(t), \tag{2.23}$$

where $h(t)$ denotes the impulse response. If a dynamic system is linear but time-variant, its impulse response depends on the time origin; if $\delta(t)$ yields $h(t)$, then $\delta(t-\tau)$ produces $h(t, \tau)$. Thus,

$$y(t) = h(t, \tau) * x(t). \tag{2.24}$$

Finally, if a system is both nonlinear and dynamic, then its impulse response can be approximated by a Volterra series. This is described in Section 2.8.

2.2 EFFECTS OF NONLINEARITY

While analog and RF circuits can be approximated by a linear model for small-signal operation, nonlinearities often lead to interesting and important phenomena that are not predicted by small-signal models. In this section, we study these phenomena for memoryless systems whose input/output characteristic can be approximated by[2]

$$y(t) \approx \alpha_1 x(t) + \alpha_2 x^2(t) + \alpha_3 x^3(t). \tag{2.25}$$

The reader is cautioned, however, that the effect of storage elements (dynamic nonlinearity) and higher-order nonlinear terms must be carefully examined to ensure (2.25) is a plausible representation. Section 2.7 deals with the case of dynamic nonlinearity. We may consider α_1 as the small-signal gain of the system because the other two terms are negligible for small input swings. For example, $\alpha_1 = -\sqrt{\mu_n C_{ox}(W/L) I_{SS}} R_D$ in Eq. (2.22).

The nonlinearity effects described in this section primarily arise from the third-order term in Eq. (2.25). The second-order term too manifests itself in certain types of receivers and is studied in Chapter 4.

2.2.1 Harmonic Distortion

If a sinusoid is applied to a nonlinear system, the output generally exhibits frequency components that are integer multiples ("harmonics") of the input frequency. In Eq. (2.25), if $x(t) = A\cos\omega t$, then

$$y(t) = \alpha_1 A\cos\omega t + \alpha_2 A^2\cos^2\omega t + \alpha_3 A^3\cos^3\omega t \tag{2.26}$$

$$= \alpha_1 A\cos\omega t + \frac{\alpha_2 A^2}{2}(1 + \cos 2\omega t) + \frac{\alpha_3 A^3}{4}(3\cos\omega t + \cos 3\omega t) \tag{2.27}$$

$$= \frac{\alpha_2 A^2}{2} + \left(\alpha_1 A + \frac{3\alpha_3 A^3}{4}\right)\cos\omega t + \frac{\alpha_2 A^2}{2}\cos 2\omega t + \frac{\alpha_3 A^3}{4}\cos 3\omega t. \tag{2.28}$$

---

2. Note that this expression should be considered as a fit across the signal swings of interest rather than as a Taylor expansion in the vicinity of $x = 0$. These two views may yield slightly different values for α_j.

In Eq. (2.28), the first term on the right-hand side is a dc quantity arising from second-order nonlinearity, the second is called the "fundamental," the third is the second harmonic, and the fourth is the third harmonic. We sometimes say that even-order nonlinearity introduces dc offsets.

From the above expansion, we make two observations. First, even-order harmonics result from α_j with even j, and vanish if the system has odd symmetry, i.e., if it is fully differential. In reality, however, random mismatches corrupt the symmetry, yielding finite even-order harmonics. Second, in (2.28) the amplitudes of the second and third harmonics are proportional to A^2 and A^3, respectively, i.e., we say the nth harmonic grows in proportion to A^n.

In many RF circuits, harmonic distortion is unimportant or an irrelevant indicator of the effect of nonlinearity. For example, an amplifier operating at 2.4 GHz produces a second harmonic at 4.8 GHz, which is greatly suppressed if the circuit has a narrow bandwidth. Nonetheless, harmonics must always be considered carefully before they are dismissed. The following examples illustrate this point.

Example 2.5

An analog multiplier "mixes" its two inputs as shown in Fig. 2.7, ideally producing $y(t) = kx_1(t)x_2(t)$, where k is a constant.[3] Assume $x_1(t) = A_1 \cos \omega_1 t$ and $x_2(t) = A_2 \cos \omega_2 t$.

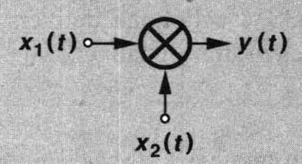


Figure 2.7 *Analog multiplier.*

(a) If the mixer is ideal, determine the output frequency components.
(b) If the input port sensing $x_2(t)$ suffers from third-order nonlinearity, determine the output frequency components.

Solution:

(a) We have

$$y(t) = k(A_1 \cos \omega_1 t)(A_2 \cos \omega_2 t) \tag{2.29}$$

$$= \frac{kA_1A_2}{2} \cos(\omega_1 + \omega_2)t + \frac{kA_1A_2}{2} \cos(\omega_1 - \omega_2)t. \tag{2.30}$$

The output thus contains the sum and difference frequencies. These may be considered "desired" components.

(Continues)

3. The factor k is necessary to ensure a proper dimension for $y(t)$.

Example 2.5 *(Continued)*

(b) Representing the third harmonic of $x_2(t)$ by $(\alpha_3 A_2^3/4)\cos 3\omega_2 t$, we write

$$y(t) = k(A_1 \cos \omega_1 t)\left(A_2 \cos \omega_2 t + \frac{\alpha_3 A_2^3}{4}\cos 3\omega_2 t\right) \tag{2.31}$$

$$= \frac{kA_1A_2}{2}\cos(\omega_1+\omega_2)t + \frac{kA_1A_2}{2}\cos(\omega_1-\omega_2)t$$
$$+ \frac{k\alpha_3 A_1 A_2^3}{8}\cos(\omega_1+3\omega_2)t + \frac{k\alpha_3 A_1 A_2^3}{8}\cos(\omega_1-3\omega_2)t. \tag{2.32}$$

The mixer now produces two "spurious" components at $\omega_1+3\omega_2$ and $\omega_1-3\omega_2$, one or both of which often prove problematic. For example, if $\omega_1 = 2\pi \times (850\text{ MHz})$ and $\omega_2 = 2\pi \times (900\text{ MHz})$, then $|\omega_1 - 3\omega_2| = 2\pi \times (1850\text{ MHz})$, an "undesired" component that is difficult to filter because it lies close to the desired component at $\omega_1 + \omega_2 = 2\pi \times (1750\text{ MHz})$.

Example 2.6

The transmitter in a 900-MHz GSM cellphone delivers 1 W of power to the antenna. Explain the effect of the harmonics of this signal.

Solution:

The second harmonic falls within another GSM cell phone band around 1800 MHz and must be sufficiently small to negligibly impact the other users in that band. The third, fourth, and fifth harmonics do not coincide with any popular bands but must still remain below a certain level imposed by regulatory organizations in each country. The sixth harmonic falls in the 5-GHz band used in wireless local area networks (WLANs), e.g., in laptops. Figure 2.8 summarizes these results.

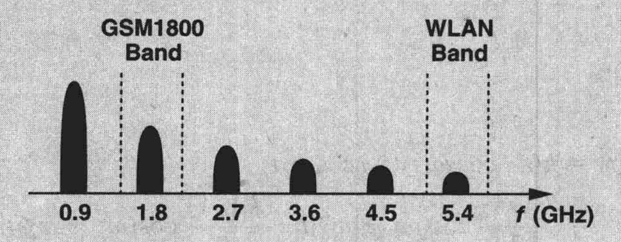


Figure 2.8 *Summary of harmonic components.*

2.2.2 Gain Compression

The small-signal gain of circuits is usually obtained with the assumption that harmonics are negligible. However, our formulation of harmonics, as expressed by Eq. (2.28), indicates

that the gain experienced by $A\cos\omega t$ is equal to $\alpha_1 + 3\alpha_3 A^2/4$ and hence varies appreciably as A becomes larger.[4] We must then ask, do α_1 and α_3 have the same sign or opposite signs? Returning to the third-order polynomial in Eq. (2.25), we note that if $\alpha_1\alpha_3 > 0$, then $\alpha_1 x + \alpha_3 x^3$ overwhelms $\alpha_2 x^2$ for large x regardless of the sign of α_2, yielding an "expansive" characteristic [Fig. 2.9(a)]. For example, an ideal bipolar transistor operating in the forward active region produces a collector current in proportion to $\exp(V_{BE}/V_T)$, exhibiting expansive behavior. On the other hand, if $\alpha_1\alpha_3 < 0$, the term $\alpha_3 x^3$ "bends" the characteristic for sufficiently large x [Fig. 2.9(b)], leading to "compressive" behavior, i.e., a decreasing gain as the input amplitude increases. For example, the differential pair of Fig. 2.6(a) suffers from compression as the second term in (2.22) becomes comparable with the first. Since most RF circuits of interest are compressive, we hereafter focus on this type.

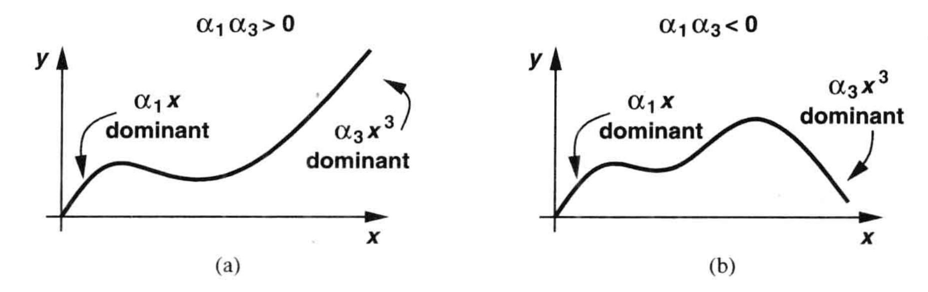


Figure 2.9 *(a) Expansive and (b) compressive characteristics.*

With $\alpha_1\alpha_3 < 0$, the gain experienced by $A\cos\omega t$ in Eq. (2.28) falls as A rises. We quantify this effect by the "1-dB compression point," defined as the input signal level that causes the gain to drop by 1 dB. If plotted on a log-log scale as a function of the input level, the output level, A_{out}, falls below its ideal value by 1 dB at the 1-dB compression point, $A_{in,1dB}$ (Fig. 2.10). Note that (a) A_{in} and A_{out} are voltage quantities here, but compression can also be expressed in terms of power quantities; (b) 1-dB compression may also be specified in terms of the output level at which it occurs, $A_{out,1dB}$. The input and output compression points typically prove relevant in the receive path and the transmit path, respectively.

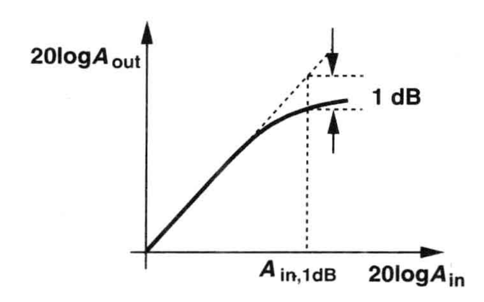


Figure 2.10 *Definition of 1-dB compression point.*

4. This effect is akin to the fact that nonlinearity can also be viewed as variation of the *slope* of the input/output characteristic with the input level.

To calculate the input 1-dB compression point, we equate the compressed gain, $\alpha_1 + (3\alpha_3/4)A^2_{in,1dB}$, to 1 dB less than the ideal gain, α_1:

$$20\log\left|\alpha_1 + \frac{3}{4}\alpha_3 A^2_{in,1dB}\right| = 20\log|\alpha_1| - 1\text{ dB}. \quad (2.33)$$

It follows that

$$A_{in,1dB} = \sqrt{0.145\left|\frac{\alpha_1}{\alpha_3}\right|}. \quad (2.34)$$

Note that Eq. (2.34) gives the *peak* value (rather than the peak-to-peak value) of the input. Also denoted by P_{1dB}, the 1-dB compression point is typically in the range of -20 to -25 dBm (63.2 to 35.6 mV_{pp} in 50-Ω system) at the input of RF receivers. We use the notations A_{1dB} and P_{1dB} interchangeably in this book. Whether they refer to the input or the output will be clear from the context or specified explicitly. While gain compression by 1 dB seems arbitrary, the 1-dB compression point represents a 10% reduction in the gain and is widely used to characterize RF circuits and systems.

Why does compression matter? After all, it appears that if a signal is so large as to reduce the gain of a receiver, then it must lie well above the receiver noise and be easily detectable. In fact, for some modulation schemes, this statement holds and compression of the receiver would seem benign. For example, as illustrated in Fig. 2.11(a), a frequency-modulated signal carries no information in its amplitude and hence tolerates compression (i.e., amplitude limiting). On the other hand, modulation schemes that contain information in the amplitude are distorted by compression [Fig. 2.11(b)]. This issue manifests itself in both receivers and transmitters.

Another adverse effect arising from compression occurs if a large *interferer* accompanies the received signal [Fig. 2.12(a)]. In the time domain, the small desired signal is superimposed on the large interferer. Consequently, the receiver gain is reduced by the large excursions produced by the interferer even though the desired signal itself is small

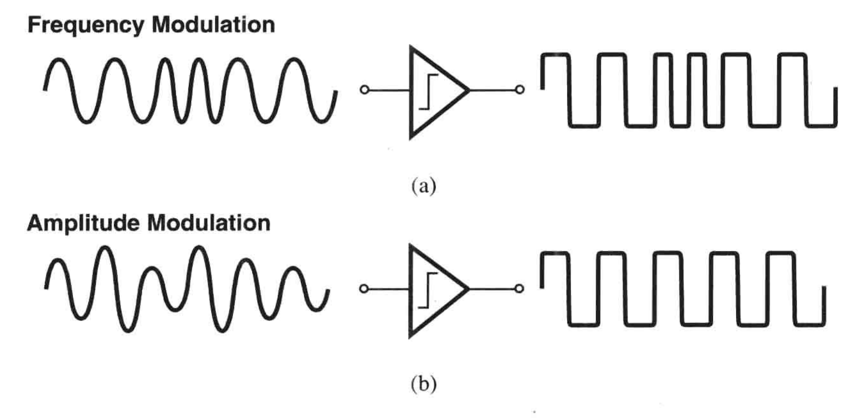


Figure 2.11 *Effect of compressive nonlinearity on (a) FM and (b) AM waveforms.*

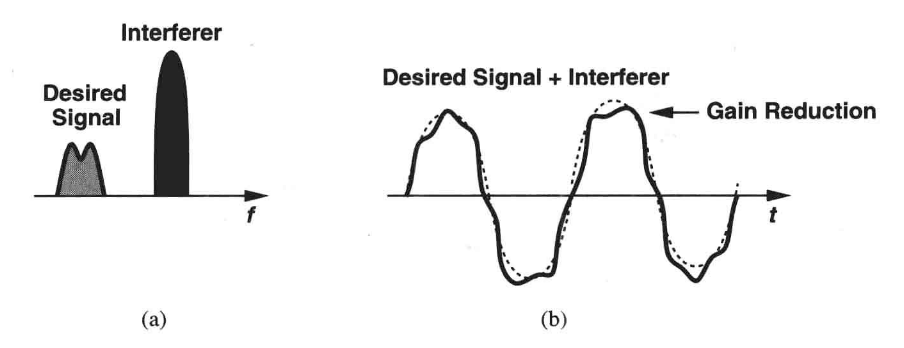


Figure 2.12 *(a) Interferer accompanying signal, (b) effect in time domain.*

[Fig. 2.12(b)]. Called "desensitization," this phenomenon lowers the signal-to-noise ratio (SNR) at the receiver output and proves critical even if the signal contains no amplitude information.

To quantify desensitization, let us assume $x(t) = A_1 \cos \omega_1 t + A_2 \cos \omega_2 t$, where the first and second terms represent the desired component and the interferer, respectively. With the third-order characteristic of Eq. (2.25), the output appears as

$$y(t) = \left(\alpha_1 + \frac{3}{4}\alpha_3 A_1^2 + \frac{3}{2}\alpha_3 A_2^2\right) A_1 \cos \omega_1 t + \cdots . \tag{2.35}$$

Note that α_2 is absent in compression. For $A_1 \ll A_2$, this reduces to

$$y(t) = \left(\alpha_1 + \frac{3}{2}\alpha_3 A_2^2\right) A_1 \cos \omega_1 t + \cdots . \tag{2.36}$$

Thus, the gain experienced by the desired signal is equal to $\alpha_1 + 3\alpha_3 A_2^2/2$, a decreasing function of A_2 if $\alpha_1 \alpha_3 < 0$. In fact, for sufficiently large A_2, the gain drops to zero, and we say the signal is "blocked." In RF design, the term "blocking signal" or "blocker" refers to interferers that desensitize a circuit even if they do not reduce the gain to zero. Some RF receivers must be able to withstand blockers that are 60 to 70 dB greater than the desired signal.

Example 2.7

A 900-MHz GSM transmitter delivers a power of 1 W to the antenna. By how much must the second harmonic of the signal be suppressed (filtered) so that it does not desensitize a 1.8-GHz receiver having $P_{1dB} = -25$ dBm? Assume the receiver is 1 m away (Fig. 2.13) and the 1.8-GHz signal is attenuated by 10 dB as it propagates across this distance.

(Continues)

Example 2.7 *(Continued)*

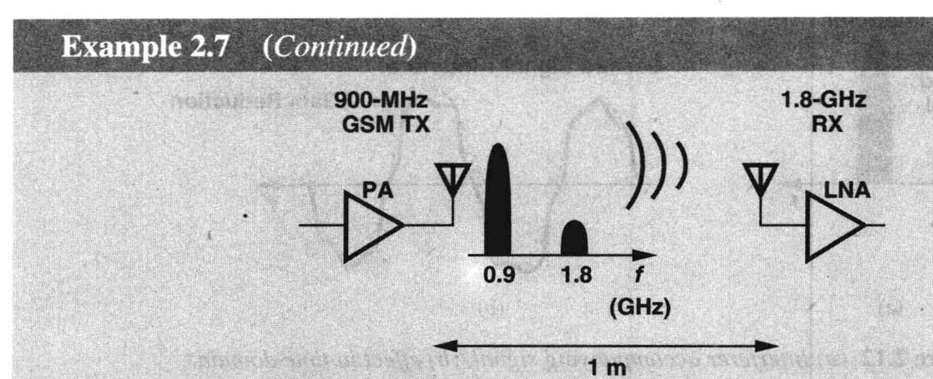


Figure 2.13 *TX and RX in a cellular system.*

Solution:

The output power at 900 MHz is equal to +30 dBm. With an attenuation of 10 dB, the second harmonic must not exceed −15 dBm at the transmitter antenna so that it is below P_{1dB} of the receiver. Thus, the second harmonic must remain at least 45 dB below the fundamental at the TX output. In practice, this interference must be another several dB lower to ensure the RX does not compress.

2.2.3 Cross Modulation

Another phenomenon that occurs when a weak signal and a strong interferer pass through a nonlinear system is the *transfer* of modulation from the interferer to the signal. Called "cross modulation," this effect is exemplified by Eq. (2.36), where variations in A_2 affect the amplitude of the signal at ω_1. For example, suppose that the interferer is an amplitude-modulated signal, $A_2(1 + m\cos\omega_m t)\cos\omega_2 t$, where m is a constant and ω_m denotes the modulating frequency. Equation (2.36) thus assumes the following form:

$$y(t) = \left[\alpha_1 + \frac{3}{2}\alpha_3 A_2^2\left(1 + \frac{m^2}{2} + \frac{m^2}{2}\cos 2\omega_m t + 2m\cos\omega_m t\right)\right] A_1\cos\omega_1 t + \cdots. \quad (2.37)$$

In other words, the desired signal at the output suffers from amplitude modulation at ω_m and $2\omega_m$. Figure 2.14 illustrates this effect.

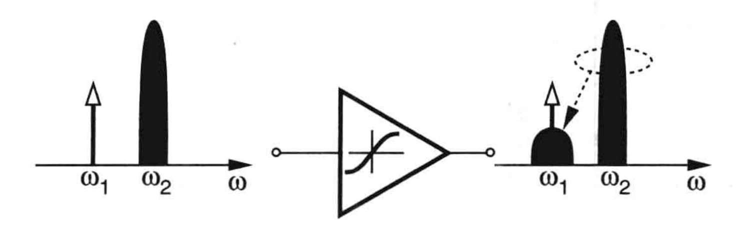


Figure 2.14 *Cross modulation.*

Example 2.8

Suppose an interferer contains phase modulation but not amplitude modulation. Does cross modulation occur in this case?

Solution:

Expressing the input as $x(t) = A_1 \cos \omega_1 t + A_2 \cos(\omega_2 t + \phi)$, where the second term represents the interferer (A_2 is constant but ϕ varies with time), we use the third-order polynomial in Eq. (2.25) to write

$$y(t) = \alpha_1[A_1 \cos \omega_1 t + A_2 \cos(\omega_2 t + \phi)] + \alpha_2[A_1 \cos \omega_1 t + A_2 \cos(\omega_2 t + \phi)]^2 + \alpha_3[A_1 \cos \omega_1 t + A_2 \cos(\omega_2 t + \phi)]^3. \quad (2.38)$$

We now note that (1) the second-order term yields components at $\omega_1 \pm \omega_2$ but not at ω_1; (2) the third-order term expansion gives $3\alpha_3 A_1 \cos \omega_1 t A_2^2 \cos^2(\omega_2 t + \phi)$, which, according to $\cos^2 x = (1 + \cos 2x)/2$, results in a component at ω_1. Thus,

$$y(t) = \left(\alpha_1 + \frac{3}{2}\alpha_3 A_2^2\right) A_1 \cos \omega_1 t + \cdots. \quad (2.39)$$

Interestingly, the desired signal at ω_1 does not experience cross modulation. That is, phase-modulated interferers do not cause cross modulation in *memoryless* (static) nonlinear systems. Dynamic nonlinear systems, on the other hand, may not follow this rule.

Cross modulation commonly arises in amplifiers that must simultaneously process many independent signal channels. Examples include cable television transmitters and systems employing "orthogonal frequency division multiplexing" (OFDM). We examine OFDM in Chapter 3.

2.2.4 Intermodulation

Our study of nonlinearity has thus far considered the case of a single signal (for harmonic distortion) or a signal accompanied by one large interferer (for desensitization). Another scenario of interest in RF design occurs if *two* interferers accompany the desired signal. Such a scenario represents realistic situations and reveals nonlinear effects that may not manifest themselves in a harmonic distortion or desensitization test.

If two interferers at ω_1 and ω_2 are applied to a nonlinear system, the output generally exhibits components that are not harmonics of these frequencies. Called "intermodulation" (IM), this phenomenon arises from "mixing" (multiplication) of the two components as their sum is raised to a power greater than unity. To understand how Eq. (2.25) leads to intermodulation, assume $x(t) = A_1 \cos \omega_1 t + A_2 \cos \omega_2 t$. Thus,

$$y(t) = \alpha_1(A_1 \cos \omega_1 t + A_2 \cos \omega_2 t) + \alpha_2(A_1 \cos \omega_1 t + A_2 \cos \omega_2 t)^2 + \alpha_3(A_1 \cos \omega_1 t + A_2 \cos \omega_2 t)^3. \quad (2.40)$$

Expanding the right-hand side and discarding the dc terms, harmonics, and components at $\omega_1 \pm \omega_2$, we obtain the following "intermodulation products":

$$\omega = 2\omega_1 \pm \omega_2 : \quad \frac{3\alpha_3 A_1^2 A_2}{4}\cos(2\omega_1 + \omega_2)t + \frac{3\alpha_3 A_1^2 A_2}{4}\cos(2\omega_1 - \omega_2)t \quad (2.41)$$

$$\omega = 2\omega_2 \pm \omega_1 : \quad \frac{3\alpha_3 A_1 A_2^2}{4}\cos(2\omega_2 + \omega_1)t + \frac{3\alpha_3 A_1 A_2^2}{4}\cos(2\omega_2 - \omega_1)t \quad (2.42)$$

and these fundamental components:

$$\omega = \omega_1,\ \omega_2 : \left(\alpha_1 A_1 + \frac{3}{4}\alpha_3 A_1^3 + \frac{3}{2}\alpha_3 A_1 A_2^2\right)\cos\omega_1 t + \left(\alpha_1 A_2 + \frac{3}{4}\alpha_3 A_2^3 + \frac{3}{2}\alpha_3 A_2 A_1^2\right)\cos\omega_2 t \quad (2.43)$$

Figure 2.15 illustrates the results. Among these, the third-order IM products at $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$ are of particular interest. This is because, if ω_1 and ω_2 are close to each other, then $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$ appear in the vicinity of ω_1 and ω_2. We now explain the significance of this statement.

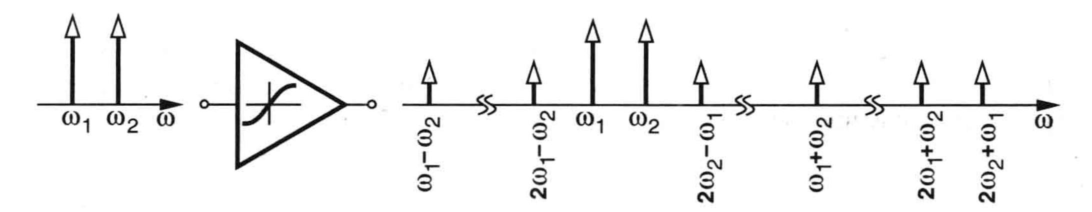


Figure 2.15 *Generation of various intermodulation components in a two-tone test.*

Suppose an antenna receives a small desired signal at ω_0 along with two large interferers at ω_1 and ω_2, providing this combination to a low-noise amplifier (Fig. 2.16). Let us assume that the interferer frequencies happen to satisfy $2\omega_1 - \omega_2 = \omega_0$. Consequently, the intermodulation product at $2\omega_1 - \omega_2$ falls onto the desired channel, corrupting the signal.

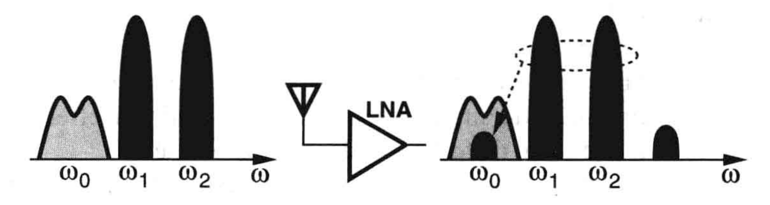


Figure 2.16 *Corruption due to third-order intermodulation.*

Example 2.9

Suppose four Bluetooth users operate in a room as shown in Fig. 2.17. User 4 is in the receive mode and attempts to sense a weak signal transmitted by User 1 at 2.410 GHz.

Example 2.9 (*Continued*)

At the same time, Users 2 and 3 transmit at 2.420 GHz and 2.430 GHz, respectively. Explain what happens.

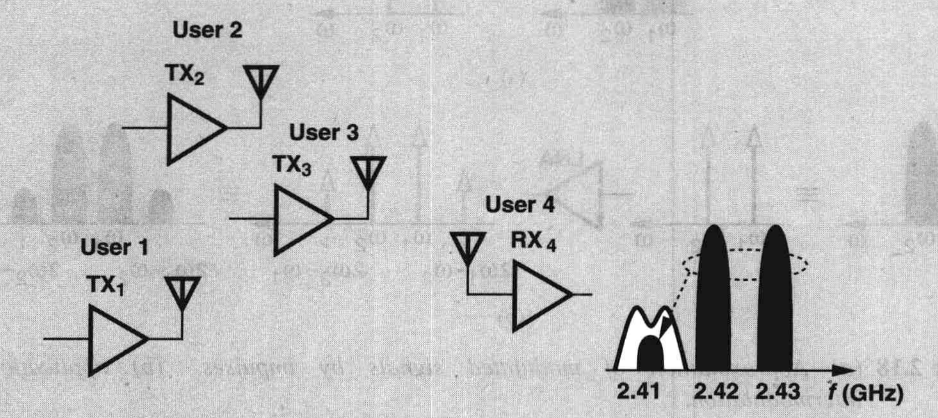


Figure 2.17 *Bluetooth RX in the presence of several transmitters.*

Solution:

Since the frequencies transmitted by Users 1, 2, and 3 happen to be equally spaced, the intermodulation in the LNA of RX_4 corrupts the desired signal at 2.410 GHz.

The reader may raise several questions at this point: (1) In our analysis of intermodulation, we represented the interferers with pure (unmodulated) sinusoids (called "tones") whereas in Figs. 2.16 and 2.17, the interferers are modulated. Are these consistent? (2) Can gain compression and desensitization (P_{1dB}) also model intermodulation, or do we need other measures of nonlinearity? (3) Why can we not simply remove the interferers by filters so that the receiver does not experience intermodulation? We answer the first two here and address the third in Chapter 4.

For narrowband signals, it is sometimes helpful to "condense" their energy into an impulse, i.e., represent them with a tone of equal power [Fig. 2.18(a)]. This approximation must be made judiciously: if applied to study gain compression, it yields reasonably accurate results; on the other hand, if applied to the case of cross modulation, it fails. In intermodulation analyses, we proceed as follows: (a) approximate the interferers with tones, (b) calculate the level of intermodulation products at the output, and (c) mentally convert the intermodulation tones back to modulated components so as to see the corruption.[5] This thought process is illustrated in Fig. 2.18(b).

We now deal with the second question: if the gain is not compressed, then can we say that intermodulation is negligible? The answer is no; the following example illustrates this point.

5. Since a tone contains no randomness, it generally does not corrupt a signal. But a tone appearing in the spectrum of a signal may make the detection difficult.

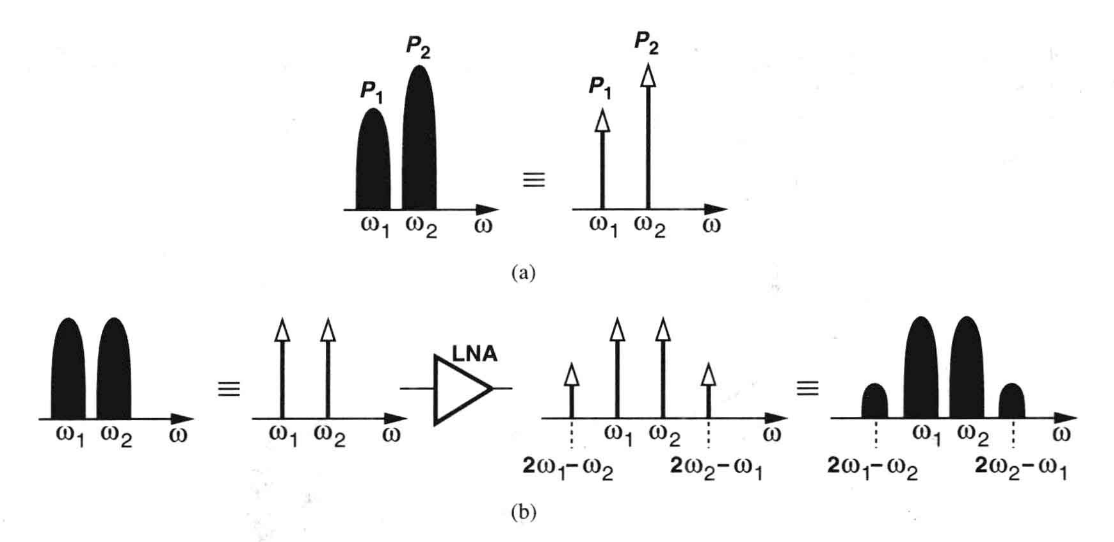


Figure 2.18 *(a) Approximation of modulated signals by impulses, (b) application to intermodulation.*

Example 2.10

A Bluetooth receiver employs a low-noise amplifier having a gain of 10 and an input impedance of 50 Ω. The LNA senses a desired signal level of −80 dBm at 2.410 GHz and two interferers of equal levels at 2.420 GHz and 2.430 GHz. For simplicity, assume the LNA drives a 50-Ω load.

(a) Determine the value of α_3 that yields a P_{1dB} of −30 dBm.

(b) If each interferer is 10 dB below P_{1dB}, determine the corruption experienced by the desired signal at the LNA output.

Solution:

(a) Noting that $-30\,\text{dBm} = 20\,\text{mV}_{\text{pp}} = 10\,\text{mV}_{\text{p}}$, from Eq. (2.34), we have $\sqrt{0.145|\alpha_1/\alpha_3|} = 10\,\text{mV}_{\text{p}}$. Since $\alpha_1 = 10$, we obtain $\alpha_3 = 14,500\,\text{V}^{-2}$.

(b) Each interferer has a level of −40 dBm (= 6.32 mV$_{\text{pp}}$). Setting $A_1 = A_2 = 6.32\,\text{mV}_{\text{pp}}/2$ in Eq. (2.41), we determine the amplitude of the IM product at 2.410 GHz as

$$\frac{3\alpha_3 A_1^2 A_2}{4} = 0.343\,\text{mV}_{\text{p}} = -59.3\,\text{dBm}. \tag{2.44}$$

The desired signal is amplified by a factor of $\alpha_1 = 10 = 20$ dB, emerging at the output at a level of −60 dBm. Unfortunately, the IM product is as large as the signal itself even though the LNA does not experience significant compression.

The two-tone test is versatile and powerful because it can be applied to systems with arbitrarily narrow bandwidths. A sufficiently small difference between the two tone frequencies ensures that the IM products also fall within the band, thereby providing a

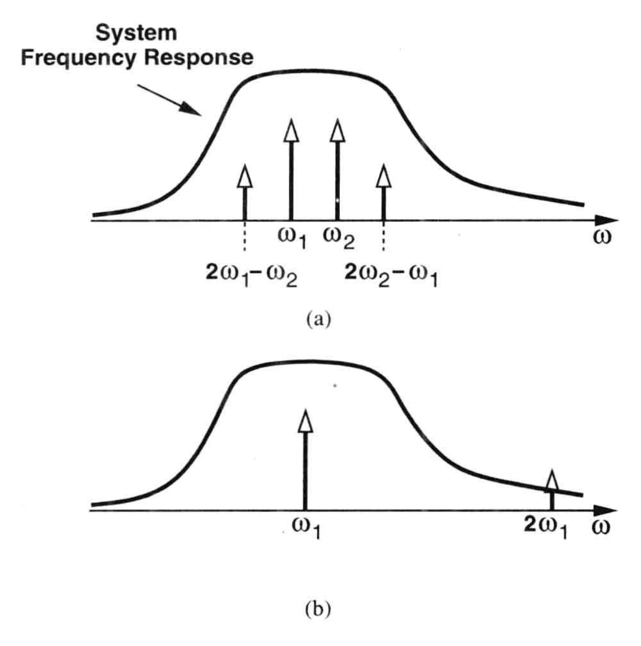


Figure 2.19 *(a) Two-tone and (b) harmonic tests in a narrowband system.*

meaningful view of the nonlinear behavior of the system. Depicted in Fig. 2.19(a), this attribute stands in contrast to harmonic distortion tests, where higher harmonics lie so far away in frequency that they are heavily filtered, making the system appear quite linear [Fig. 2.19(b)].

Third Intercept Point Our thoughts thus far indicate the need for a measure of intermodulation. A common method of IM characterization is the "two-tone" test, whereby two pure sinusoids of *equal* amplitudes are applied to the input. The amplitude of the output IM products is then normalized to that of the fundamentals at the output. Denoting the peak amplitude of each tone by A, we can write the result as

$$\text{Relative IM} = 20 \log \left(\frac{3}{4} \frac{\alpha_3}{\alpha_1} A^2 \right) \text{ dBc}, \tag{2.45}$$

where the unit dBc denotes decibels with respect to the "carrier" to emphasize the normalization. Note that, if the amplitude of each input tone increases by 6 dB (a factor of two), the amplitude of the IM products ($\propto A^3$) rises by 18 dB and hence the *relative* IM by 12 dB.[6]

The principal difficulty in specifying the relative IM for a circuit is that it is meaningful only if the value of A is given. From a practical point of view, we prefer a *single* measure that captures the intermodulation behavior of the circuit with no need to know the input level at which the two-tone test is carried out. Fortunately, such a measure exists and is called the "third intercept point" (IP_3).

The concept of IP_3 originates from our earlier observation that, if the amplitude of each tone rises, that of the output IM products increases more sharply ($\propto A^3$). Thus, if we continue to raise A, the amplitude of the IM products eventually becomes *equal* to that

6. It is assumed that no compression occurs so that the output fundamental tones also rise by 6 dB.

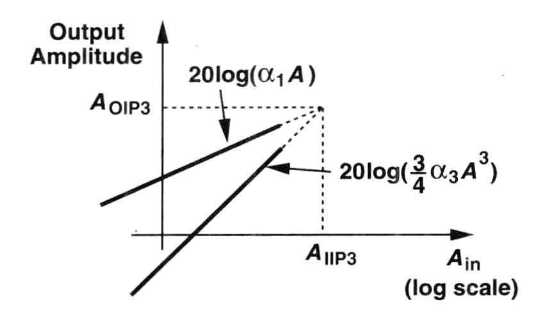


Figure 2.20 *Definition of IP_3 (for voltage quantities).*

of the fundamental tones at the output. As illustrated in Fig. 2.20 on a log-log scale, the input level at which this occurs is called the "input third intercept point" (IIP_3). Similarly, the corresponding output is represented by OIP_3. In subsequent derivations, we denote the input amplitude as A_{IIP3}.

To determine the IIP_3, we simply equate the fundamental and IM amplitudes:

$$|\alpha_1 A_{IIP3}| = \left|\frac{3}{4}\alpha_3 A_{IIP3}^3\right|, \tag{2.46}$$

obtaining

$$A_{IIP3} = \sqrt{\frac{4}{3}\left|\frac{\alpha_1}{\alpha_3}\right|}. \tag{2.47}$$

Interestingly,

$$\frac{A_{IIP3}}{A_{1dB}} = \sqrt{\frac{4}{0.435}} \tag{2.48}$$

$$\approx 9.6\,\text{dB}. \tag{2.49}$$

This ratio proves helpful as a sanity check in simulations and measurements.[7] We sometimes write IP_3 rather than IIP_3 if it is clear from the context that the input is of interest.

Upon further consideration, the reader may question the consistency of the above derivations. If the IP_3 is 9.6 dB *higher* than P_{1dB}, is the gain not heavily compressed at $A_{in} = A_{IIP3}$?! If the gain is compressed, why do we still express the amplitude of the fundamentals at the output as $\alpha_1 A$? It appears that we must instead write this amplitude as $[\alpha_1 + (9/4)\alpha_3 A^2]A$ to account for the compression.

In reality, the situation is even more complicated. The value of IP_3 given by Eq. (2.47) may *exceed* the supply voltage, indicating that higher-order nonlinearities manifest themselves as A_{in} approaches A_{IIP3} [Fig. 2.21(a)]. In other words, the IP_3 is not a directly measureable quantity.

In order to avoid these quandaries, we measure the IP_3 as follows. We begin with a very low input level so that $\alpha_1 + (9/4)\alpha_3 A_{in}^2 \approx \alpha_1$ (and, of course, higher order nonlinearities

7. Note that this relationship holds for a third-order system and not necessarily if higher-order terms manifest themselves.

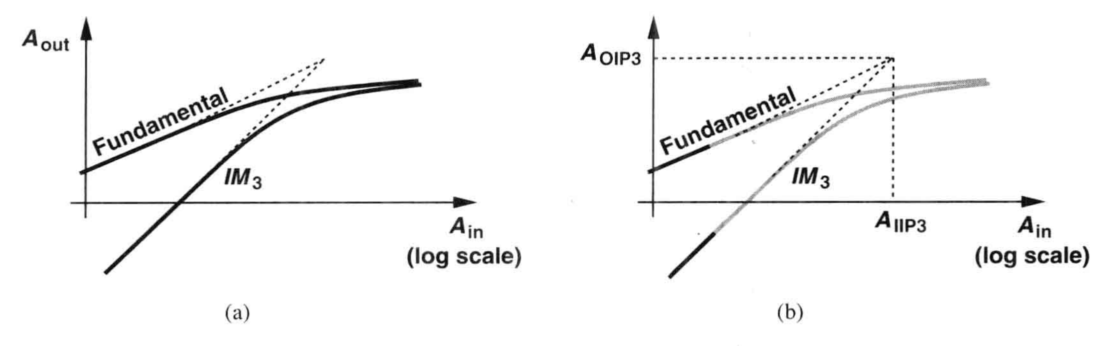


Figure 2.21 *(a) Actual behavior of nonlinear circuits, (b) definition of IP_3 based on extrapolation.*

are also negligible). We increase A_{in}, plot the amplitudes of the fundamentals and the IM products on a log-log scale, and *extrapolate* these plots according to their slopes (one and three, respectively) to obtain the IP_3 [Fig. 2.21(b)]. To ensure that the signal levels remain well below compression and higher-order terms are negligible, we must observe a 3-dB rise in the IM products for every 1-dB increase in A_{in}. On the other hand, if A_{in} is excessively small, then the output IM components become comparable with the noise floor of the circuit (or the noise floor of the simulated spectrum), thus leading to inaccurate results.

Example 2.11

A low-noise amplifier senses a −80-dBm signal at 2.410 GHz and two −20-dBm interferers at 2.420 GHz and 2.430 GHz. What IIP_3 is required if the IM products must remain 20 dB below the signal? For simplicity, assume 50-Ω interfaces at the input and output.

Solution:

Denoting the peak amplitudes of the signal and the interferers by A_{sig} and A_{int}, respectively, we can write at the LNA output:

$$20 \log |\alpha_1 A_{sig}| - 20\,\text{dB} = 20 \log \left|\frac{3}{4}\alpha_3 A_{int}^3\right|. \tag{2.50}$$

It follows that

$$|\alpha_1 A_{sig}| = \left|\frac{30}{4}\alpha_3 A_{int}^3\right|. \tag{2.51}$$

In a 50-Ω system, the −80-dBm and −20-dBm levels respectively yield $A_{sig} = 31.6\ \mu V_p$ and $A_{int} = 31.6\ mV_p$. Thus,

$$IIP_3 = \sqrt{\frac{4}{3}\left|\frac{\alpha_1}{\alpha_3}\right|} \tag{2.52}$$

$$= 3.65\ V_p \tag{2.53}$$

$$= +15.2\,\text{dBm}. \tag{2.54}$$

Such an IP_3 is extremely difficult to achieve, especially for a complete receiver chain.

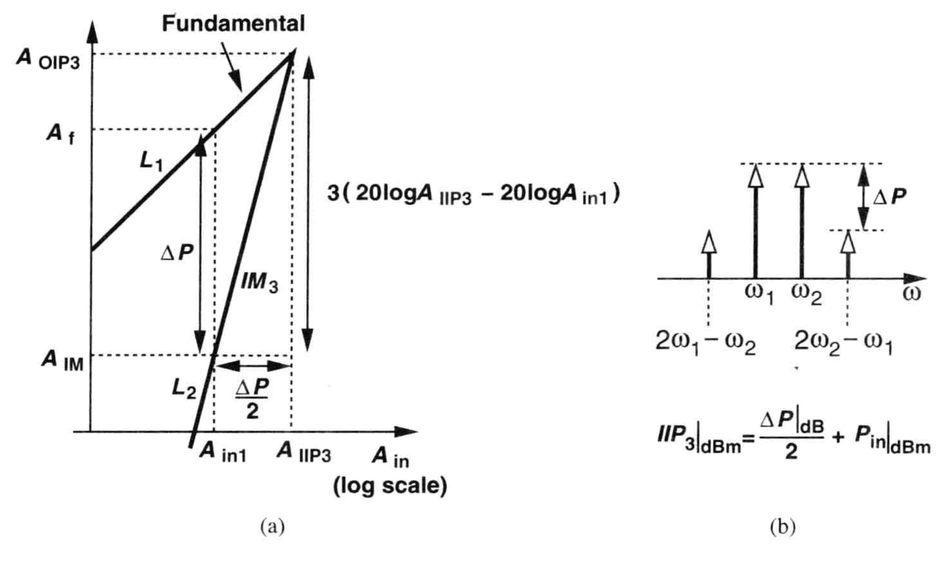


Figure 2.22 *(a) Relationships among various power levels in a two-tone test, (b) illustration of shortcut technique.*

Since extrapolation proves quite tedious in simulations or measurements, we often employ a shortcut that provides a reasonable initial estimate. As illustrated in Fig. 2.22(a), suppose *hypothetically* that the input is equal to A_{IIP3}, and hence the (extrapolated) output IM products are as large as the (extrapolated) fundamental tones. Now, the input is reduced to a level A_{in1}. That is, the change in the input is equal to $20\log A_{IIP3} - 20\log A_{in1}$. On a log-log scale, the IM products fall with a slope of 3 and the fundamentals with a slope of unity. Thus, the *difference* between the two plots increases with a slope of 2. We denote $20\log A_f - 20\log A_{IM}$ by ΔP and write

$$\Delta P = 20\log A_f - 20\log A_{IM} = 2(20\log A_{IIP3} - 20\log A_{in1}), \tag{2.55}$$

obtaining

$$20\log A_{IIP3} = \frac{\Delta P}{2} + 20\log A_{in1}. \tag{2.56}$$

In other words, for a given input level (well below P_{1dB}), the IIP_3 can be calculated by halving the difference between the output fundamental and IM levels and adding the result to the input level, where all values are expressed as logarithmic quantities. Figure 2.22(b) depicts an abbreviated notation for this rule. The key point here is that the IP_3 is measured without extrapolation.

Why do we consider the above result an *estimate*? After all, the derivation assumes third-order nonlinearity. A difficulty arises if the circuit contains *dynamic* nonlinearities, in which case this result may deviate from that obtained by extrapolation. The latter is the standard and accepted method for measuring and reporting the IP_3, but the shortcut method proves useful in understanding the behavior of the device under test.

We should remark that *second-order* nonlinearity also leads to a certain type of intermodulation and is characterized by a "second intercept point," (IP_2).[8] We elaborate on this effect in Chapter 4.

2.2.5 Cascaded Nonlinear Stages

Since in RF systems, signals are processed by cascaded stages, it is important to know how the nonlinearity of each stage is referred to the input of the cascade. The calculation of P_{1dB} for a cascade is outlined in Problem 2.1. Here, we determine the IP_3 of a cascade. For the sake of brevity, we hereafter denote the input IP_3 by A_{IP3} unless otherwise noted.

Consider two nonlinear stages in cascade (Fig. 2.23). If the input/output characteristics of the two stages are expressed, respectively, as

$$y_1(t) = \alpha_1 x(t) + \alpha_2 x^2(t) + \alpha_3 x^3(t) \tag{2.57}$$

$$y_2(t) = \beta_1 y_1(t) + \beta_2 y_1^2(t) + \beta_3 y_1^3(t), \tag{2.58}$$

then

$$y_2(t) = \beta_1[\alpha_1 x(t) + \alpha_2 x^2(t) + \alpha_3 x^3(t)] + \beta_2[\alpha_1 x(t) + \alpha_2 x^2(t) + \alpha_3 x^3(t)]^2 + \beta_3[\alpha_1 x(t) + \alpha_2 x^2(t) + \alpha_3 x^3(t)]^3. \tag{2.59}$$

Considering only the first- and third-order terms, we have

$$y_2(t) = \alpha_1\beta_1 x(t) + (\alpha_3\beta_1 + 2\alpha_1\alpha_2\beta_2 + \alpha_1^3\beta_3)x^3(t) + \cdots. \tag{2.60}$$

Thus, from Eq. (2.47),

$$A_{IP3} = \sqrt{\frac{4}{3}\left|\frac{\alpha_1\beta_1}{\alpha_3\beta_1 + 2\alpha_1\alpha_2\beta_2 + \alpha_1^3\beta_3}\right|}. \tag{2.61}$$

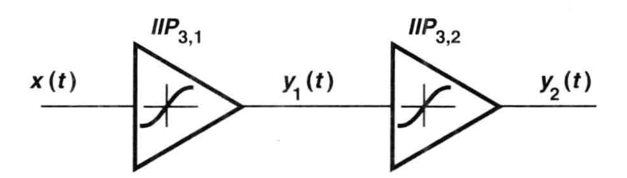


Figure 2.23 *Cascaded nonlinear stages.*

Example 2.12

Two differential pairs are cascaded. Is it possible to select the denominator of Eq. (2.61) such that IP_3 goes to infinity?

(Continues)

8. As seen in the next section, second-order nonlinearity also affects the IP_3 in cascaded systems.

Example 2.12 *(Continued)*

Solution:

With no asymmetries in the cascade, $\alpha_2 = \beta_2 = 0$. Thus, we seek the condition $\alpha_3\beta_1 + \alpha_1^3\beta_3 = 0$, or equivalently,

$$\frac{\alpha_3}{\alpha_1} = -\frac{\beta_3}{\beta_1} \cdot \alpha_1^2. \tag{2.62}$$

Since both stages are compressive, $\alpha_3/\alpha_1 < 0$ and $\beta_3/\beta_1 < 0$. It is therefore impossible to achieve an arbitrarily high IP_3.

Equation (2.61) leads to more intuitive results if its two sides are squared and inverted:

$$\frac{1}{A_{IP3}^2} = \frac{3}{4}\left|\frac{\alpha_3\beta_1 + 2\alpha_1\alpha_2\beta_2 + \alpha_1^3\beta_3}{\alpha_1\beta_1}\right| \tag{2.63}$$

$$= \frac{3}{4}\left|\frac{\alpha_3}{\alpha_1} + \frac{2\alpha_2\beta_2}{\beta_1} + \frac{\alpha_1^2\beta_3}{\beta_1}\right| \tag{2.64}$$

$$= \left|\frac{1}{A_{IP3,1}^2} + \frac{3\alpha_2\beta_2}{2\beta_1} + \frac{\alpha_1^2}{A_{IP3,2}^2}\right|, \tag{2.65}$$

where $A_{IP3,1}$ and $A_{IP3,2}$ represent the input IP_3's of the first and second stages, respectively. Note that A_{IP3}, $A_{IP3,1}$, and $A_{IP3,2}$ are voltage quantities.

The key observation in Eq. (2.65) is that to "refer" the IP_3 of the second stage to the input of the cascade, we must divide it by α_1. Thus, the higher the gain of the first stage, the more nonlinearity is contributed by the second stage.

IM Spectra in a Cascade To gain more insight into the above results, let us assume $x(t) = A\cos\omega_1 t + A\cos\omega_2 t$ and identify the IM products in a cascade. With the aid of Fig. 2.24, we make the following observations:[9]

1. The input tones are amplified by a factor of approximately α_1 in the first stage and β_1 in the second. Thus, the output fundamentals are given by $\alpha_1\beta_1 A(\cos\omega_1 t + \cos\omega_2 t)$.
2. The IM products generated by the first stage, namely, $(3\alpha_3/4)A^3[\cos(2\omega_1 - \omega_2)t + \cos(2\omega_2 - \omega_1)t]$, are amplified by a factor of β_1 when they appear at the output of the second stage.
3. Sensing $\alpha_1 A(\cos\omega_1 t + \cos\omega_2 t)$ at its input, the second stage produces its own IM components: $(3\beta_3/4)(\alpha_1 A)^3\cos(2\omega_1 - \omega_2)t + (3\beta_3/4)(\alpha_1 A)^3\cos(2\omega_2 - \omega_1)t$.

9. The spectrum of $A\cos\omega t$ consists of two impulses, each with a weight of $A/2$. We drop the factor of $1/2$ in the figures for simplicity.

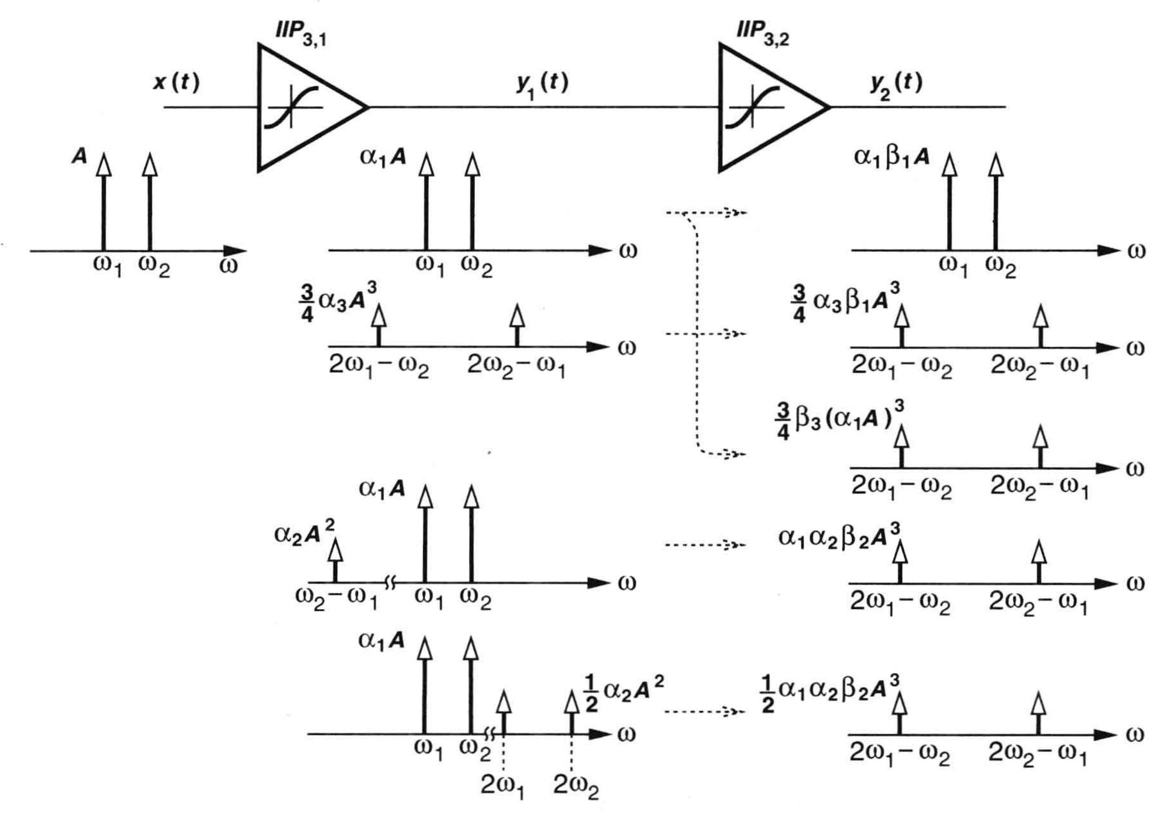


Figure 2.24 *Spectra in a cascade of nonlinear stages.*

4. The second-order nonlinearity in $y_1(t)$ generates components at $\omega_1 - \omega_2$, $2\omega_1$, and $2\omega_2$. Upon experiencing a similar nonlinearity in the second stage, these components are mixed with those at ω_1 and ω_2 and translated to $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$. Specifically, as shown in Fig. 2.24, $y_2(t)$ contains terms such as $2\beta_2[\alpha_1 A\cos\omega_1 t \times \alpha_2 A^2\cos(\omega_1 - \omega_2)t]$ and $2\beta_2(\alpha_1 A\cos\omega_1 t \times 0.5\alpha_2 A^2\cos 2\omega_2 t)$. The resulting IM products can be expressed as $(3\alpha_1\alpha_2\beta_2 A^3/2)[\cos(2\omega_1 - \omega_2)t + \cos(2\omega_2 - \omega_1)t]$. Interestingly, the cascade of two *second-order* nonlinearities can produce *third-order* IM products.

Adding the amplitudes of the IM products, we have

$$\begin{aligned} y_2(t) &= \alpha_1\beta_1 A(\cos\omega_1 t + \cos\omega_2 t) \\ &+ \left(\frac{3\alpha_3\beta_1}{4} + \frac{3\alpha_1^3\beta_3}{4} + \frac{3\alpha_1\alpha_2\beta_2}{2}\right)A^3[\cos(\omega_1 - 2\omega_2)t \\ &+ \cos(2\omega_2 - \omega_1)t] + \cdots, \end{aligned} \tag{2.66}$$

obtaining the same IP_3 as above. This result assumes zero phase shift for all components.

Why did we add the amplitudes of the IM_3 products in Eq. (2.66) without regard for their phases? Is it possible that phase shifts in the first and second stages allow partial

cancellation of these terms and hence a higher IP_3? Yes, it is possible but uncommon in practice. Since the frequencies ω_1, ω_2, $2\omega_1 - \omega_2$, and $2\omega_2 - \omega_1$ are close to one another, these components experience approximately equal phase shifts.

But how about the terms described in the fourth observation? Components such as $\omega_1 - \omega_2$ and $2\omega_1$ may fall well out of the signal band and experience phase shifts different from those in the first three observations. For this reason, we may consider Eqs. (2.65) and (2.66) as the worst-case scenario. Since most RF systems incorporate narrowband circuits, the terms at $\omega_1 \pm \omega_2$, $2\omega_1$, and $2\omega_2$ are heavily attenuated at the output of the first stage. Consequently, the second term on the right-hand side of (2.65) becomes negligible, and

$$\frac{1}{A_{IP3}^2} \approx \frac{1}{A_{IP3,1}^2} + \frac{\alpha_1^2}{A_{IP3,2}^2}. \tag{2.67}$$

Extending this result to three or more stages, we have

$$\frac{1}{A_{IP3}^2} \approx \frac{1}{A_{IP3,1}^2} + \frac{\alpha_1^2}{A_{IP3,2}^2} + \frac{\alpha_1^2\beta_1^2}{A_{IP3,3}^2} + \cdots. \tag{2.68}$$

Thus, if each stage in a cascade has a gain greater than unity, the nonlinearity of the latter stages becomes increasingly more critical because the IP_3 of each stage is equivalently scaled *down* by the total gain preceding that stage.

Example 2.13

A low-noise amplifier having an input IP_3 of $-10\,\text{dBm}$ and a gain of $20\,\text{dB}$ is followed by a mixer with an input IP_3 of $+4\,\text{dBm}$. Which stage limits the IP_3 of the cascade more? Assume the conceptual picture shown in Fig. 2.1(b) to go between volts and dBm's.

Solution:

With $\alpha_1 = 20\,\text{dB}$, we note that

$$A_{IP3,1} = -10\,\text{dBm} \tag{2.69}$$

$$\frac{A_{IP3,2}}{\alpha_1} = -16\,\text{dBm}. \tag{2.70}$$

Since the scaled IP_3 of the second stage is lower than the IP_3 of the first stage, we say the second stage limits the overall IP_3 more.

In the simulation of a cascade, it is possible to determine which stage limits the linearity more. As depicted in Fig. 2.25, we examine the relative IM magnitudes at the output of each stage (Δ_1 and Δ_2, expressed in dB.) If $\Delta_2 \approx \Delta_1$, the second stage contributes negligible nonlinearity. On the other hand, if Δ_2 is substantially less than Δ_1, then the second stage limits the IP_3.

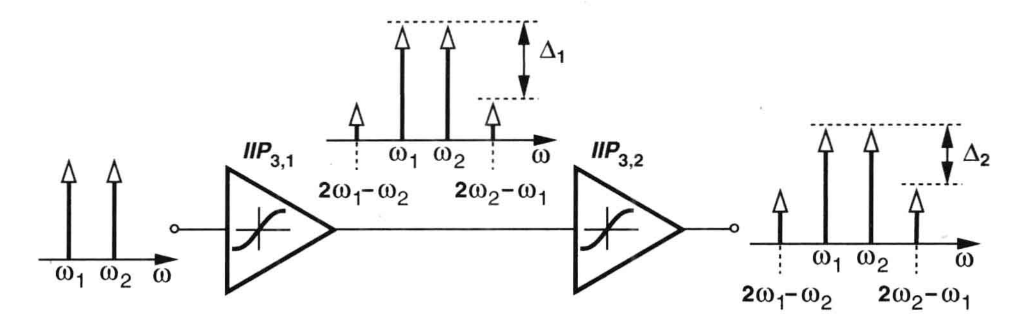


Figure 2.25 *Growth of IM components along the cascade.*

2.2.6 AM/PM Conversion

In some RF circuits, e.g., power amplifiers, amplitude modulation (AM) may be converted to phase modulation (PM), thus producing undesirable effects. In this section, we study this phenomenon.

AM/PM conversion (APC) can be viewed as the dependence of the phase shift upon the signal amplitude. That is, for an input $V_{in}(t) = V_1 \cos \omega_1 t$, the fundamental output component is given by

$$V_{out}(t) = V_2 \cos[\omega_1 t + \phi(V_1)], \tag{2.71}$$

where $\phi(V_1)$ denotes the amplitude-dependent phase shift. This, of course, does not occur in a linear time-invariant system. For example, the phase shift experienced by a sinusoid of frequency ω_1 through a first-order low-pass RC section is given by $-\tan^{-1}(RC\omega_1)$ regardless of the amplitude. Moreover, APC does not appear in a memoryless nonlinear system because the phase shift is zero in this case.

We may therefore surmise that AM/PM conversion arises if a system is both dynamic and nonlinear. For example, if the capacitor in a first-order low-pass RC section is nonlinear, then its "average" value may depend on V_1, resulting in a phase shift, $-\tan^{-1}(RC\omega_1)$, that itself varies with V_1. To explore this point, let us consider the arrangement shown in Fig. 2.26 and assume

$$C_1 = (1 + \alpha V_{out})C_0. \tag{2.72}$$

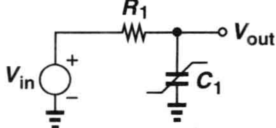


Figure 2.26 *RC section with nonlinear capacitor.*

This capacitor is considered nonlinear because its value depends on its voltage. An exact calculation of the phase shift is difficult here as it requires that we write $V_{in} = R_1 C_1 dV_{out}/dt + V_{out}$ and hence solve

$$V_1 \cos \omega_1 t = R_1(1 + \alpha V_{out})C_0 \frac{dV_{out}}{dt} + V_{out}. \tag{2.73}$$

We therefore make an approximation. Since the value of C_1 varies *periodically* with time, we can express the output as that of a first-order network but with a time-varying

capacitance, $C_1(t)$:

$$V_{out}(t) \approx \frac{V_1}{\sqrt{1 + R_1^2 C_1^2(t)\omega_1^2}} \cos\{\omega_1 t - \tan^{-1}[R_1 C_1(t)\omega_1]\}. \tag{2.74}$$

If $R_1 C_1(t)\omega_1 \ll 1$ rad,

$$V_{out}(t) \approx V_1 \cos[\omega_1 t - R_1(1 + \alpha V_{out})C_0\omega_1]. \tag{2.75}$$

We also assume that $(1 + \alpha V_{out})C_0 \approx (1 + \alpha V_1 \cos \omega_1 t)C_0$, obtaining

$$V_{out}(t) \approx V_1 \cos(\omega_1 t - R_1 C_0 \omega_1 - \alpha R_1 C_0 \omega_1 V_1 \cos \omega_1 t). \tag{2.76}$$

Does the output *fundamental* contain an input-dependent phase shift here? No, it does not! The reader can show that the third term inside the parentheses produces only higher *harmonics*. Thus, the phase shift of the fundamental is equal to $-R_1 C_0 \omega_1$ and hence constant.

The above example entails no AM/PM conversion because of the *first-order* dependence of C_1 upon V_{out}. As illustrated in Fig. 2.27, the average value of C_1 is equal to C_0 regardless of the output amplitude. In general, since C_1 varies periodically, it can be expressed as a Fourier series with a "dc" term representing its average value:

$$C_1(t) = C_{avg} + \sum_{n=1}^{\infty} a_n \cos(n\omega_1 t) + \sum_{n=1}^{\infty} b_n \sin(n\omega_1 t). \tag{2.77}$$

Thus, if C_{avg} is a function of the amplitude, then the phase shift of the fundamental component in the output voltage becomes input-dependent. The following example illustrates this point.

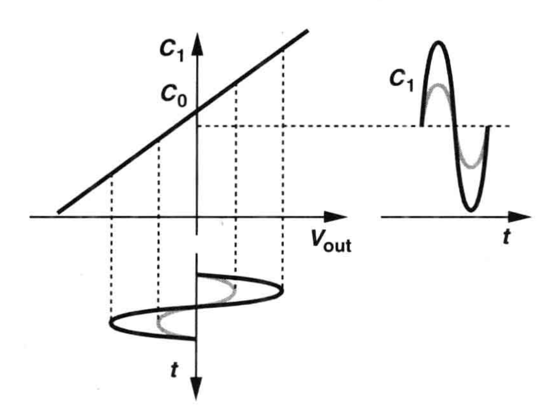


Figure 2.27 *Time variation of capacitor with first-order voltage dependence for small and large swings.*

Example 2.14

Suppose C_1 in Fig. 2.26 is expressed as $C_1 = C_0(1 + \alpha_1 V_{out} + \alpha_2 V_{out}^2)$. Study the AM/PM conversion in this case if $V_{in}(t) = V_1 \cos \omega_1 t$.

Solution:

Figure 2.28 plots $C_1(t)$ for small and large input swings, revealing that C_{avg} indeed depends on the amplitude. We rewrite Eq. (2.75) as

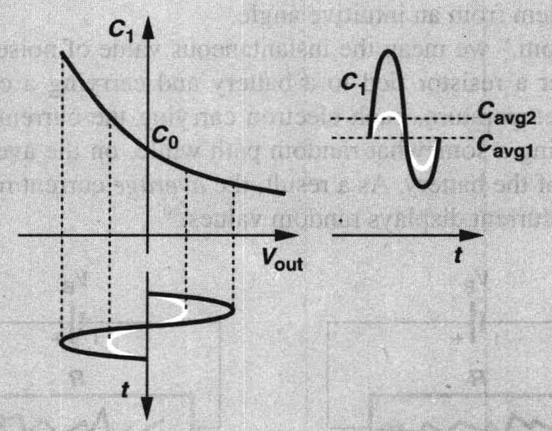


Figure 2.28 *Time variation of capacitor with second-order voltage dependence for small and large swings.*

$$V_{out}(t) \approx V_1 \cos[\omega_1 t - R_1 C_0 \omega_1 (1 + \alpha_1 V_1 \cos \omega_1 t + \alpha_2 V_1^2 \cos^2 \omega_1 t)] \tag{2.78}$$

$$\approx V_1 \cos(\omega_1 t - R_1 C_0 \omega_1 - \frac{\alpha_2 R_1 C_0 \omega_1 V_1^2}{2} - \cdots). \tag{2.79}$$

The phase shift of the fundamental now contains an input-dependent term, $-(\alpha_2 R_1 C_0 \omega_1 V_1^2)/2$. Figure 2.28 also suggests that AM/PM conversion does not occur if the capacitor voltage dependence is odd-symmetric.

What is the effect of APC? In the presence of APC, amplitude modulation (or amplitude noise) corrupts the phase of the signal. For example, if $V_{in}(t) = V_1(1 + m \cos \omega_m t) \cos \omega_1 t$, then Eq. (2.79) yields a phase corruption equal to $-\alpha_2 R_1 C_0 \omega_1 (2mV_1 \cos \omega_m t + m^2 V_1^2 \cos^2 \omega_m t)/2$. We will encounter examples of APC in Chapters 8 and 12.

2.3 NOISE

The performance of RF systems is limited by noise. Without noise, an RF receiver would be able to detect arbitrarily small inputs, allowing communication across arbitrarily long

distances. In this section, we review basic properties of noise and methods of calculating noise in circuits. For a more complete study of noise in analog circuits, the reader is referred to [1].

2.3.1 Noise as a Random Process

The trouble with noise is that it is random. Engineers who are used to dealing with well-defined, deterministic, "hard" facts often find the concept of randomness difficult to grasp, especially if it must be incorporated mathematically. To overcome this fear of randomness, we approach the problem from an intuitive angle.

By "noise is random," we mean the instantaneous value of noise cannot be predicted. For example, consider a resistor tied to a battery and carrying a current [Fig. 2.29(a)]. Due to the ambient temperature, each electron carrying the current experiences thermal agitation, thus following a somewhat random path while, on the average, moving toward the positive terminal of the battery. As a result, the *average* current remains equal to V_B/R but the instantaneous current displays random values.[10]

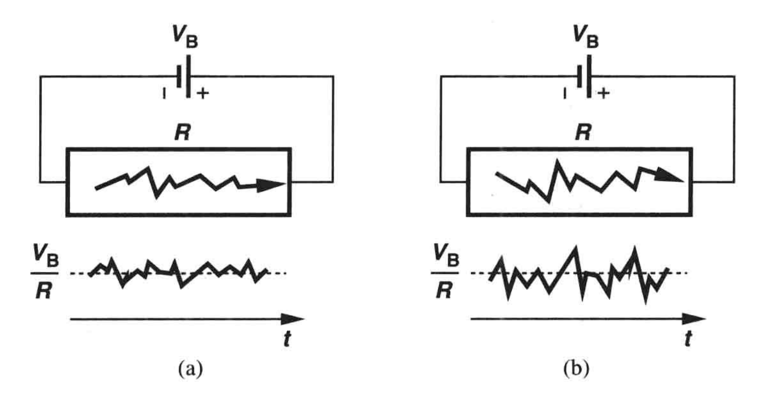


Figure 2.29 *(a) Noise generated in a resistor, (b) effect of higher temperature.*

Since noise cannot be characterized in terms of instantaneous voltages or currents, we seek other attributes of noise that are predictable. For example, we know that a higher ambient temperature leads to greater thermal agitation of electrons and hence larger fluctuations in the current [Fig. 2.29(b)]. How do we express the concept of larger random swings for a current or voltage quantity? This property is revealed by the *average power* of the noise, defined, in analogy with periodic signals, as

$$P_n = \lim_{T\to\infty} \frac{1}{T} \int_0^T n^2(t)dt, \tag{2.80}$$

where $n(t)$ represents the noise waveform. Illustrated in Fig. 2.30, this definition simply means that we compute the area under $n^2(t)$ for a long time, T, and normalize the result to T, thus obtaining the average power. For example, the two scenarios depicted in Fig. 2.29 yield different average powers.

10. As explained later, this is true even with a zero average current.

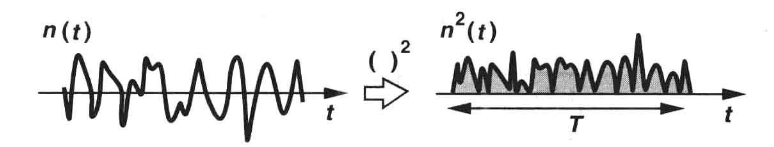


Figure 2.30 *Computation of noise power.*

If $n(t)$ is random, how do we know that P_n is not?! We are fortunate that noise components in circuits have a constant average power. For example, P_n is known and constant for a resistor at a constant ambient temperature.

How long should T in Eq. (2.80) be? Due to its randomness, noise consists of different frequencies. Thus, T must be long enough to accommodate several cycles of the *lowest* frequency. For example, the noise in a crowded restaurant arises from human voice and covers the range of 20 Hz to 20 kHz, requiring that T be on the order of 0.5 s to capture about 10 cycles of the 20-Hz components.[11]

2.3.2 Noise Spectrum

Our foregoing study suggests that the time-domain view of noise provides limited information, e.g., the average power. The frequency-domain view, on the other hand, yields much greater insight and proves more useful in RF design.

The reader may already have some intuitive understanding of the concept of "spectrum." We say the spectrum of human voice spans the range of 20 Hz to 20 kHz. This means that if we somehow measure the frequency content of the voice, we observe all components from 20 Hz to 20 kHz. How, then, do we measure a signal's frequency content, e.g., the strength of a component at 10 kHz? We would need to filter out the remainder of the spectrum and measure the *average power* of the 10-kHz component. Figure 2.31(a) conceptually illustrates such an experiment, where the microphone signal is applied to a band-pass filter having a 1-Hz bandwidth centered around 10 kHz. If a person speaks into the microphone at a steady volume, the power meter reads a constant value.

The scheme shown in Fig. 2.31(a) can be readily extended so as to measure the strength of all frequency components. As depicted in Fig. 2.31(b), a bank of 1-Hz band-pass filters centered at $f_1 \cdots f_n$ measures the average power at each frequency.[12] Called the spectrum or the "power spectral density" (PSD) of $x(t)$ and denoted by $S_x(f)$, the resulting plot displays the average power that the voice (or the noise) carries in a 1-Hz bandwidth at different frequencies.[13]

It is interesting to note that the total area under $S_x(f)$ represents the average power carried by $x(t)$:

$$\int_0^\infty S_x(f)df = \lim_{T\to\infty} \frac{1}{T}\int_0^T x^2(t)dt. \tag{2.81}$$

11. In practice, we make a guess for T, calculate P_n, increase T, recalculate P_n, and repeat until consecutive values of P_n become nearly equal.

12. This is also the conceptual operation of spectrum analyzers.

13. In the theory of signals and systems, the PSD is defined as the Fourier transform of the autocorrelation of a signal. These two views are equivalent.

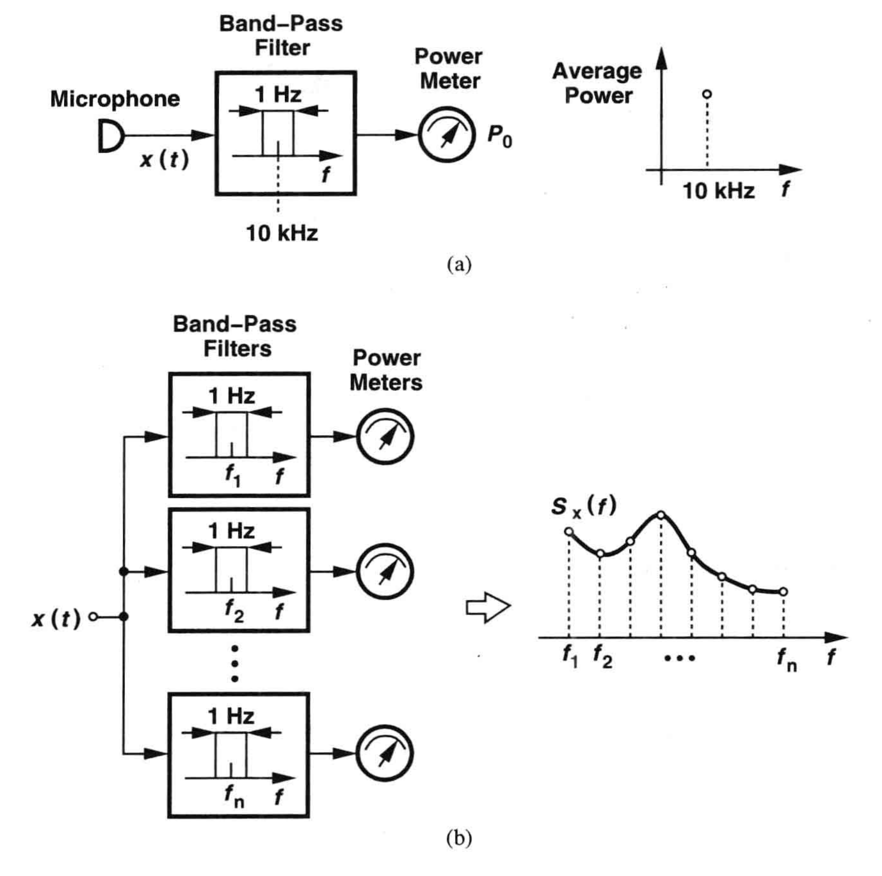


Figure 2.31 *Measurement of (a) power in 1 Hz, and (b) the spectrum.*

The spectrum shown in Fig. 2.31(b) is called "one-sided" because it is constructed for positive frequencies. In some cases, the analysis is simpler if a "two-sided" spectrum is utilized. The latter is an even-symmetric of the former scaled down vertically by a factor of two (Fig. 2.32), so that the two carry equal energies.

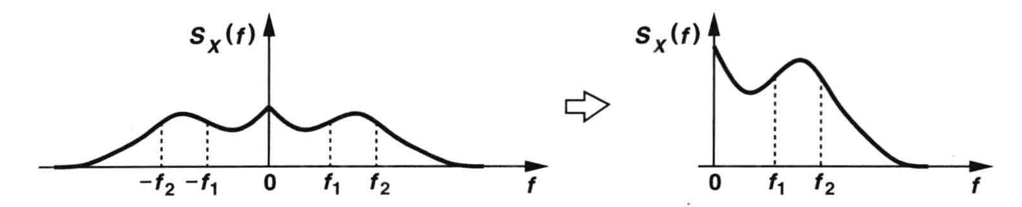


Figure 2.32 *Two-sided and one-sided spectra.*

Example 2.15

A resistor of value R_1 generates a noise voltage whose one-sided PSD is given by

$$S_v(f) = 4kTR_1, \tag{2.82}$$

Example 2.15 *(Continued)*

where $k = 1.38 \times 10^{-23}$ J/K denotes the Boltzmann constant and T the absolute temperature. Such a flat PSD is called "white" because, like white light, it contains all frequencies with equal power levels.

(a) What is the total average power carried by the noise voltage?

(b) What is the dimension of $S_v(f)$?

(c) Calculate the noise voltage for a 50-Ω resistor in 1 Hz at room temperature.

Solution:

(a) The area under $S_v(f)$ appears to be infinite, an implausible result because the resistor noise arises from the finite ambient heat. In reality, $S_v(f)$ begins to fall at $f > 1$ THz, exhibiting a finite total energy, i.e., thermal noise is not quite white.

(b) The dimension of $S_v(f)$ is voltage squared per unit bandwidth (V^2/Hz) rather than power per unit bandwidth (W/Hz). In fact, we may write the PSD as

$$\overline{V_n^2} = 4kTR, \tag{2.83}$$

where $\overline{V_n^2}$ denotes the average power of V_n in 1 Hz.[14] While some texts express the right-hand side as $4kTR\Delta f$ to indicate the total noise in a bandwidth of Δf, we omit Δf with the understanding that our PSDs always represent power in 1 Hz. We shall use $S_v(f)$ and $\overline{V_n^2}$ interchangeably.

(c) For a 50-Ω resistor at $T = 300$ K,

$$\overline{V_n^2} = 8.28 \times 10^{-19}\ \mathrm{V^2/Hz}. \tag{2.84}$$

This means that if the noise voltage of the resistor is applied to a 1-Hz band-pass filter centered at any frequency (< 1 THz), then the average measured output is given by the above value. To express the result as a root-mean-squared (rms) quantity and in more familiar units, we may take the square root of both sides:

$$\sqrt{\overline{V_n^2}} = 0.91\ \mathrm{nV}/\sqrt{\mathrm{Hz}}. \tag{2.85}$$

The familiar unit is nV but the strange unit is $\sqrt{\mathrm{Hz}}$. The latter bears no profound meaning; it simply says that the average power in 1 Hz is $(0.91\ \mathrm{nV})^2$.

2.3.3 Effect of Transfer Function on Noise

The principal reason for defining the PSD is that it allows many of the frequency-domain operations used with deterministic signals to be applied to random signals as well. For

14. Also called "spot noise."

example, if white noise is applied to a low-pass filter, how do we determine the PSD at the output? As shown in Fig. 2.33, we intuitively expect that the output PSD assumes the shape of the filter's frequency response. In fact, if $x(t)$ is applied to a linear, time-invariant system with a transfer function $H(s)$, then the output spectrum is

$$S_y(f) = S_x(f)|H(f)|^2, \tag{2.86}$$

where $H(f) = H(s = j2\pi f)$ [2]. We note that $|H(f)|$ is squared because $S_x(f)$ is a (voltage or current) squared quantity.

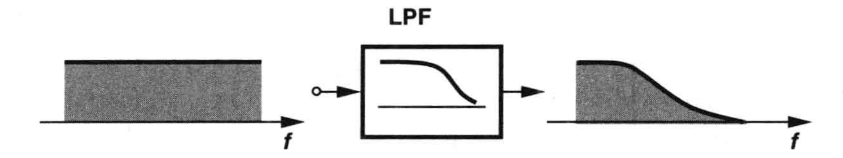


Figure 2.33 *Effect of low-pass filter on white noise.*

2.3.4 Device Noise

In order to analyze the noise performance of circuits, we wish to model the noise of their constituent elements by familiar components such as voltage and current sources. Such a representation allows the use of standard circuit analysis techniques.

Thermal Noise of Resistors As mentioned previously, the ambient thermal energy leads to random agitation of charge carriers in resistors and hence noise. The noise can be modeled by a series voltage source with a PSD of $\overline{V_n^2} = 4kTR_1$ [Thevenin equivalent, Fig. 2.34(a)] or a parallel current source with a PSD of $\overline{I_n^2} = \overline{V_n^2}/R_1 = 4kT/R_1$ [Norton equivalent, Fig. 2.34(b)]. The choice of the model sometimes simplifies the analysis. The polarity of the sources is unimportant (but must be kept the same throughout the calculations of a given circuit).

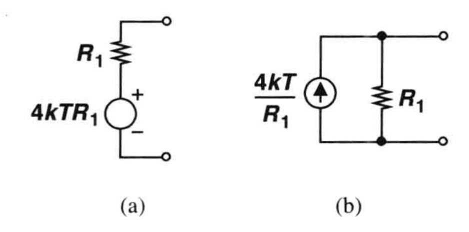


Figure 2.34 *(a) Thevenin and (b) Norton models of resistor thermal noise.*

Example 2.16

Sketch the PSD of the noise voltage measured across the parallel RLC tank depicted in Fig. 2.35(a).

Example 2.16 *(Continued)*

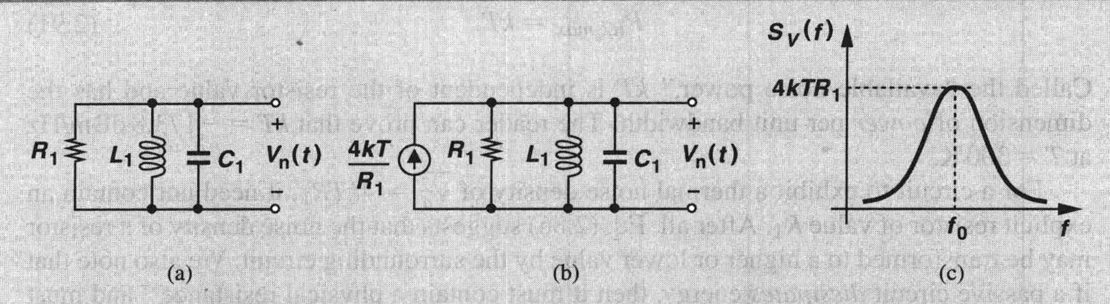


Figure 2.35 *(a) RLC tank, (b) inclusion of resistor noise, (c) output noise spectrum due to R_1.*

Solution:

Modeling the noise of R_1 by a current source, $\overline{I_{n1}^2} = 4kT/R_1$, [Fig. 2.35(b)] and noting that the transfer function V_n/I_{n1} is, in fact, equal to the impedance of the tank, Z_T, we write from Eq. (2.86)

$$\overline{V_n^2} = \overline{I_{n1}^2}|Z_T|^2. \tag{2.87}$$

At $f_0 = (2\pi\sqrt{L_1C_1})^{-1}$, L_1 and C_1 resonate, reducing the circuit to only R_1. Thus, the output noise at f_0 is simply equal to $\overline{I_{n1}^2}R_1^2 = 4kTR_1$. At lower or higher frequencies, the impedance of the tank falls and so does the output noise [Fig. 2.35(c)].

If a resistor converts the ambient heat to a noise voltage or current, can we extract energy from the resistor? In particular, does the arrangement shown in Fig. 2.36 deliver energy to R_2? Interestingly, if R_1 and R_2 reside at the same temperature, no net energy is transferred between them because R_2 also produces a noise PSD of $4kTR_2$ (Problem 2.8). However, suppose R_2 is held at $T = 0$ K. Then, R_1 continues to draw thermal energy from its environment, converting it to noise and delivering the energy to R_2. The average power transferred to R_2 is equal to

$$P_{R2} = \frac{\overline{V_{out}^2}}{R_2} \tag{2.88}$$

$$= \overline{V_n^2}\left(\frac{R_2}{R_1 + R_2}\right)^2 \frac{1}{R_2} \tag{2.89}$$

$$= 4kT\frac{R_1R_2}{(R_1 + R_2)^2}. \tag{2.90}$$

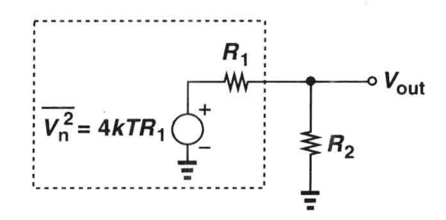


Figure 2.36 *Transfer of noise from one resistor to another.*

This quantity reaches a maximum if $R_2 = R_1$:

$$P_{R2,max} = kT. \tag{2.91}$$

Called the "available noise power," kT is independent of the resistor value and has the dimension of *power* per unit bandwidth. The reader can prove that $kT = -173.8$ dBm/Hz at $T = 300$ K.

For a circuit to exhibit a thermal noise density of $\overline{V_n^2} = 4kTR_1$, it need not contain an explicit resistor of value R_1. After all, Eq. (2.86) suggests that the noise density of a resistor may be transformed to a higher or lower value by the surrounding circuit. We also note that if a passive circuit *dissipates* energy, then it must contain a physical resistance[15] and must therefore *produce* thermal noise. We loosely say "lossy circuits are noisy."

A theorem that consolidates the above observations is as follows: If the real part of the impedance seen between two terminals of a passive (reciprocal) network is equal to $Re\{Z_{out}\}$, then the PSD of the thermal noise seen between these terminals is given by $\overline{V_n^2} = 4kTRe\{Z_{out}\}$ (Fig. 2.37) [8]. This general theorem is not limited to lumped circuits. For example, consider a transmitting antenna that dissipates energy by radiation according to the equation $V_{TX,rms}^2/R_{rad}$, where R_{rad} is the "radiation resistance" [Fig. 2.38(a)]. As a receiving element [Fig. 2.38(b)], the antenna *generates* a thermal noise PSD of[16]

$$\overline{V_{n,ant}^2} = 4kTR_{rad}. \tag{2.92}$$

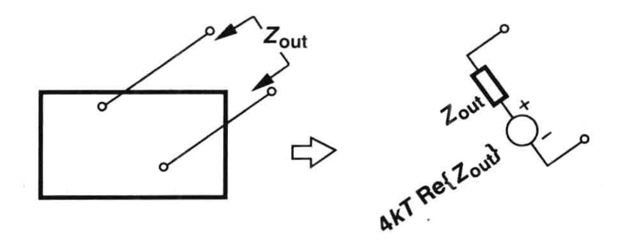


Figure 2.37 *Output noise of a passive (reciprocal) circuit.*

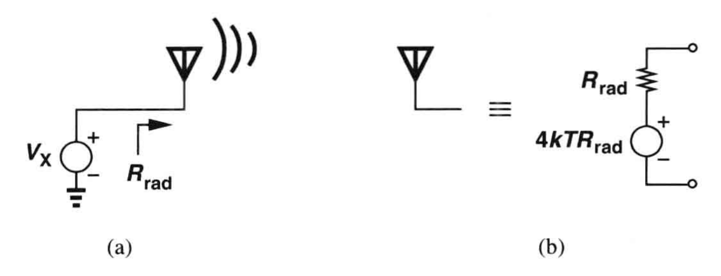


Figure 2.38 *(a) Transmitting antenna, (b) receiving antenna producing thermal noise.*

15. Recall that ideal inductors and capacitors *store* energy but do not dissipate it.

16. Strictly speaking, this is not correct because the noise of a receiving antenna is in fact given by the "background" noise (e.g., cosmic radiation). However, in RF design, the antenna noise is commonly assumed to be $4kTR_{rad}$.

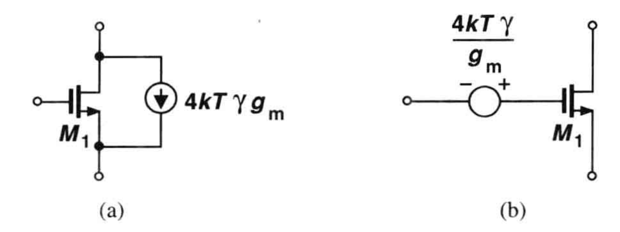


Figure 2.39 *Thermal channel noise of a MOSFET modeled as a (a) current source, (b) voltage source.*

Noise in MOSFETs The thermal noise of MOS transistors operating in the saturation region is approximated by a current source tied between the source and drain terminals [Fig. 2.39(a)]:

$$\overline{I_n^2} = 4kT\gamma g_m, \tag{2.93}$$

where γ is the "excess noise coefficient" and g_m the transconductance.[17] The value of γ is 2/3 for long-channel transistors and may rise to even 2 in short-channel devices [4]. The actual value of γ has other dependencies [5] and is usually obtained by measurements for each generation of CMOS technology. In Problem 2.10, we prove that the noise can alternatively be modeled by a voltage source $\overline{V_n^2} = 4kT\gamma/g_m$ in series with the gate [Fig. 2.39(b)].

Another component of thermal noise arises from the gate resistance of MOSFETs, an effect that becomes increasingly more important as the gate length is scaled down. Illustrated in Fig. 2.40(a) for a device with a width of W and a length of L, this resistance amounts to

$$R_G = \frac{W}{L} R_{\square}, \tag{2.94}$$

where $R_{\square}$ denotes the sheet resistance (resistance of one square) of the polysilicon gate. For example, if $W = 1\ \mu$m, $L = 45$ nm, and $R_{\square} = 15\ \Omega$, then $R_G = 333\ \Omega$. Since R_G is *distributed* over the width of the transistor [Fig. 2.40(b)], its noise must be calculated carefully. As proved in [6], the structure can be reduced to a lumped model having an equivalent gate resistance of $R_G/3$ with a thermal noise PSD of $4kTR_G/3$ [Fig. 2.40(c)]. In a good design, this noise must be much less than that of the channel:

$$4kT\frac{R_G}{3} \ll \frac{4kT\gamma}{g_m}. \tag{2.95}$$

The gate and drain terminals also exhibit physical resistances, which are minimized through the use of multiple fingers.

At very high frequencies the thermal noise current flowing through the channel couples to the gate capacitively, thus generating a "gate-induced noise current" [3] (Fig. 2.41). This

17. More accurately, $\overline{I_n^2} = 4kT\gamma g_{d0}$, where g_{d0} is the drain-source conductance in the triode region (even though the noise is measured in saturation) [3].

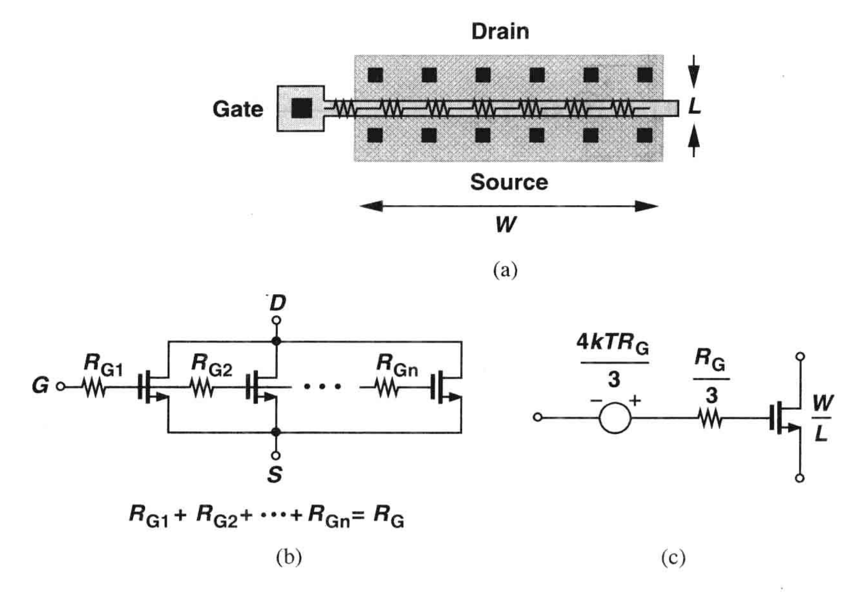


Figure 2.40 *(a) Gate resistance of a MOSFET, (b) equivalent circuit for noise calculation, (c) equivalent noise and resistance in lumped model.*

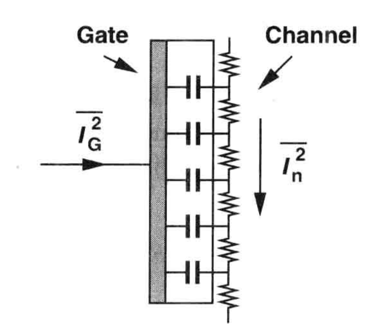


Figure 2.41 *Gate-induced noise,* $\overline{I_G^2}$.

effect is not modeled in typical circuit simulators, but its significance has remained unclear. In this book, we neglect the gate-induced noise current.

MOS devices also suffer from "flicker" or "$1/f$" noise. Modeled by a voltage source in series with the gate, this noise exhibits the following PSD:

$$\overline{V_n^2} = \frac{K}{WLC_{ox}}\frac{1}{f}, \tag{2.96}$$

where K is a process-dependent constant. In most CMOS technologies, K is lower for PMOS devices than for NMOS transistors because the former carry charge well below the silicon-oxide interface and hence suffer less from "surface states" (dangling bonds) [1]. The $1/f$ dependence means that noise components that vary slowly assume a large amplitude. The choice of the lowest frequency in the noise integration depends on the time scale of interest and/or the spectrum of the desired signal [1].

Example 2.17

Can the flicker noise be modeled by a current source?

Solution:

Yes, as shown in Fig. 2.42, a MOSFET having a small-signal voltage source of magnitude V_1 in series with its gate is equivalent to a device with a current source of value $g_m V_1$ tied between drain and source. Thus,

$$\overline{I_1^2} = g_m^2 \frac{K}{WLC_{ox}} \frac{1}{f}. \tag{2.97}$$

Figure 2.42 *Conversion of flicker noise voltage to current.*

For a given device size and bias current, the $1/f$ noise PSD intercepts the thermal noise PSD at some frequency, called the "$1/f$ noise corner frequency," f_c. Illustrated in Fig. 2.43, f_c can be obtained by converting the flicker noise voltage to current (according to the above example) and equating the result to the thermal noise current:

$$\frac{K}{WLC_{ox}} \frac{1}{f_c} g_m^2 = 4kT\gamma g_m. \tag{2.98}$$

It follows that

$$f_c = \frac{K}{WLC_{ox}} \frac{g_m}{4kT\gamma}. \tag{2.99}$$

The corner frequency falls in the range of tens or even hundreds of megahertz in today's MOS technologies.

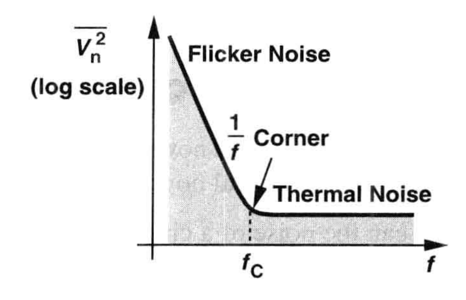


Figure 2.43 *Flicker noise corner frequency.*

While the effect of flicker noise may seem negligible at high frequencies, we must note that nonlinearity or time variance in circuits such as mixers and oscillators may translate the $1/f$-shaped spectrum to the RF range. We study these phenomena in Chapters 6 and 8.

Noise in Bipolar Transistors Bipolar transistors contain physical resistances in their base, emitter, and collector regions, all of which generate thermal noise. Moreover, they also suffer from "shot noise" associated with the transport of carriers across the base-emitter junction. As shown in Fig. 2.44, this noise is modeled by two current sources having the following PSDs:

$$\overline{I_{n,b}^2} = 2qI_B = 2q\frac{I_C}{\beta} \tag{2.100}$$

$$\overline{I_{n,c}^2} = 2qI_C, \tag{2.101}$$

where I_B and I_C are the base and collector bias currents, respectively. Since $g_m = I_C/(kT/q)$ for bipolar transistors, the collector current shot noise is often expressed as

$$\overline{I_{n,c}^2} = 4kT\frac{g_m}{2}, \tag{2.102}$$

in analogy with the thermal noise of MOSFETs or resistors.

In low-noise circuits, the base resistance thermal noise and the collector current shot noise become dominant. For this reason, wide transistors biased at high current levels are employed.

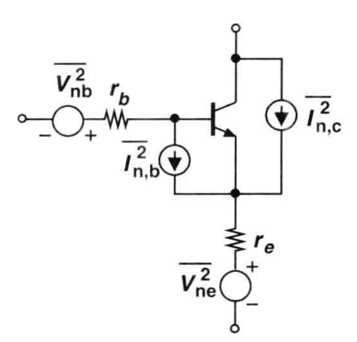


Figure 2.44 *Noise sources in a bipolar transistor.*

2.3.5 Representation of Noise in Circuits

With the noise of devices formulated above, we now wish to develop *measures* of the noise performance of circuits, i.e., metrics that reveal how noisy a given circuit is.

Input-Referred Noise How can the noise of a circuit be observed in the laboratory? We have access only to the output and hence can measure only the output noise. Unfortunately, the output noise does not permit a fair comparison between circuits: a circuit may exhibit high output noise because it has a high gain rather than high noise. For this reason, we "refer" the noise to the input.

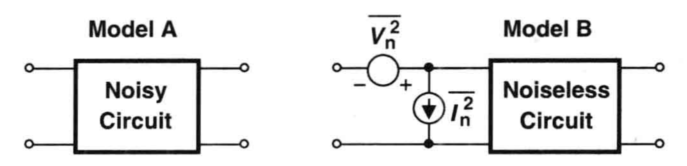


Figure 2.45 *Input-referred noise.*

In analog design, the input-referred noise is modeled by a series voltage source and a parallel current source (Fig. 2.45) [1]. The former is obtained by shorting the input port of models A and B and equating their output noises (or, equivalently, dividing the output noise by the voltage gain). Similarly, the latter is computed by leaving the input ports open and equating the output noises (or, equivalently, dividing the output noise by the transimpedance gain).

Example 2.18

Calculate the input-referred noise of the common-gate stage depicted in Fig. 2.46(a). Assume I_1 is ideal and neglect the noise of R_1.

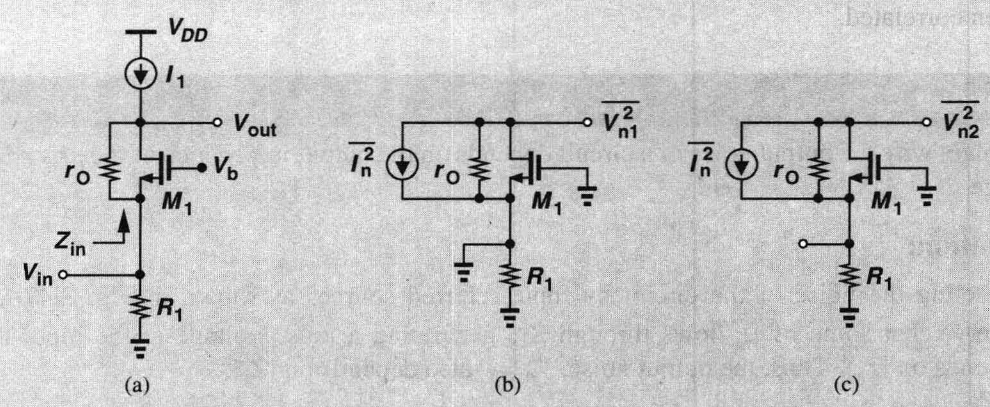


Figure 2.46 *(a) CG stage, (b) computation of input-referred noise voltage, (c) computation of input-referred noise current.*

Solution:

Shorting the input to ground, we write from Fig. 2.46(b),

$$\overline{V_{n1}^2} = \overline{I_n^2} \cdot r_O^2. \tag{2.103}$$

Since the voltage gain of the stage is given by $1 + g_m r_O$, the input-referred noise voltage is equal to

$$\overline{V_{n,in}^2} = \frac{\overline{I_n^2} r_O^2}{(1 + g_m r_O)^2} \tag{2.104}$$

$$\approx \frac{4kT\gamma}{g_m}, \tag{2.105}$$

(Continues)

Example 2.18 *(Continued)*

where it is assumed $g_m r_O \gg 1$. Leaving the input open as shown in Fig. 2.46(c), the reader can show that (Problem 2.12)

$$\overline{V_{n2}^2} = \overline{I_n^2} r_O^2. \tag{2.106}$$

Defined as the output voltage divided by the input current, the transimpedance gain of the stage is given by $g_m r_O R_1$ (why?). It follows that

$$\overline{I_{n,in}^2} = \frac{\overline{I_n^2} r_O^2}{g_m^2 r_O^2 R_1^2} \tag{2.107}$$

$$= \frac{4kT\gamma}{g_m R_1^2}. \tag{2.108}$$

From the above example, it may appear that the noise of M_1 is "counted" twice. It can be shown that [1] the two input-referred noise sources are necessary and sufficient, but often correlated.

Example 2.19

Explain why the output noise of a circuit depends on the output impedance of the *preceding* stage.

Solution:

Modeling the noise of the circuit by input-referred sources as shown in Fig. 2.47, we observe that some of $\overline{I_n^2}$ flows through Z_1, generating a noise voltage at the input that depends on $|Z_1|$. Thus, the output noise, $V_{n,out}$, also depends on $|Z_1|$.

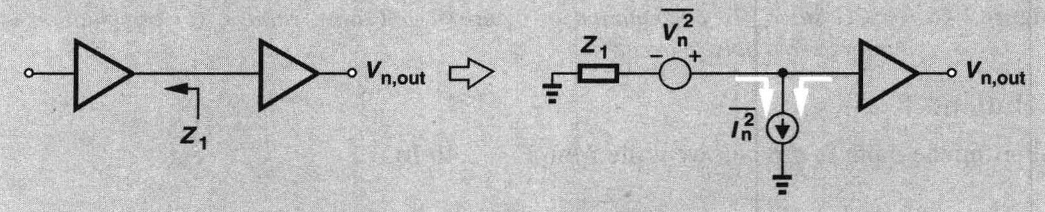


Figure 2.47 *Noise in a cascade.*

The computation and use of input-referred noise sources prove difficult at high frequencies. For example, it is quite challenging to measure the transimpedance gain of an RF stage. For this reason, RF designers employ the concept of "noise figure" as another metric of noise performance that more easily lends itself to measurement.

Noise Figure In circuit and system design, we are interested in the signal-to-noise ratio (SNR), defined as the signal power divided by the noise power. It is therefore helpful to

ask, how does the SNR degrade as the signal travels through a given circuit? If the circuit contains no noise, then the output SNR is *equal* to the input SNR even if the circuit acts as an attenuator.[18] To quantify how noisy the circuit is, we define its noise figure (NF) as

$$\text{NF} = \frac{SNR_{in}}{SNR_{out}} \tag{2.109}$$

such that it is equal to 1 for a noiseless stage. Since each quantity in this ratio has a dimension of power (or voltage squared), we express NF in decibels as

$$\text{NF}|_{\text{dB}} = 10 \log \frac{SNR_{in}}{SNR_{out}}. \tag{2.110}$$

Note that most texts call (2.109) the "noise factor" and (2.110) the noise figure. We do not make this distinction in this book.

Compared to input-referred noise, the definition of NF in (2.109) may appear rather complicated: it depends on not only the noise of the circuit under consideration but the SNR provided by the *preceding stage*. In fact, if the input signal contains no noise, then $SNR_{in} = \infty$ and $\text{NF} = \infty$, even though the circuit may have finite internal noise. For such a case, NF is not a meaningful parameter and only the input-referred noise can be specified.

Calculation of the noise figure is generally simpler than Eq. (2.109) may suggest. For example, suppose a low-noise amplifier senses the signal received by an antenna [Fig. 2.48(a)]. As predicted by Eq. (2.92), the antenna "radiation resistance," R_S, produces thermal noise, leading to the model shown in Fig. 2.48(b). Here, $\overline{V_{n,RS}^2}$ represents the thermal noise of the antenna, and $\overline{V_n^2}$ the output noise of the LNA. We must compute SNR_{in} at the LNA input and SNR_{out} at its output.

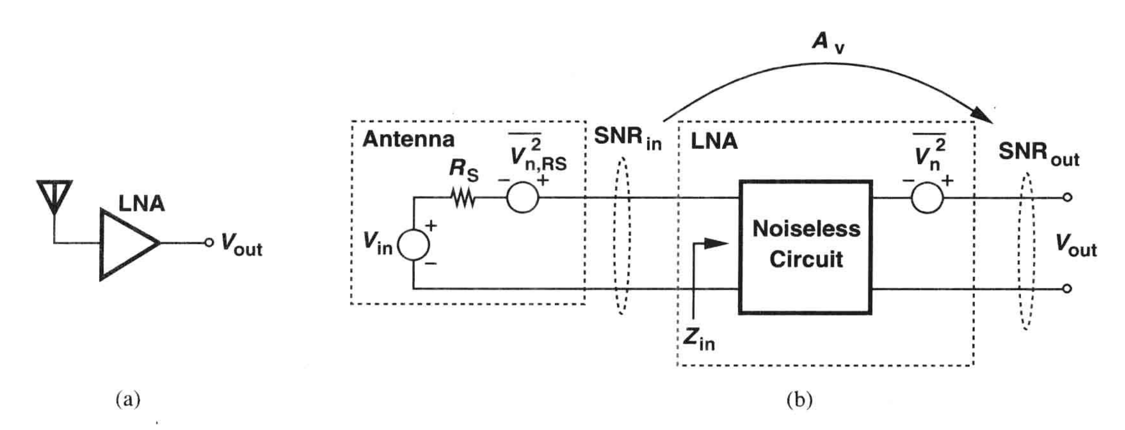


Figure 2.48 *(a) Antenna followed by LNA, (b) equivalent circuit.*

18. Because the input signal and the input noise are attenuated by the same factor.

If the LNA exhibits an input impedance of Z_{in}, then both V_{in} and V_{RS} experience an attenuation factor of $\alpha = Z_{in}/(Z_{in} + R_S)$ as they appear at the input of the LNA. That is,

$$SNR_{in} = \frac{|\alpha|^2 V_{in}^2}{|\alpha|^2 \overline{V_{RS}^2}}, \tag{2.111}$$

where V_{in} denotes the rms value of the signal received by the antenna.

To determine SNR_{out}, we assume a voltage gain of A_v from the LNA input to the output and recognize that the output signal power is equal to $V_{in}^2|\alpha|^2 A_v^2$. The output noise consists of two components: (a) the noise of the antenna amplified by the LNA, $\overline{V_{RS}^2}|\alpha|^2 A_v^2$, and (b) the output noise of the LNA, $\overline{V_n^2}$. Since these two components are uncorrelated, we simply add the PSDs and write

$$SNR_{out} = \frac{V_{in}^2 |\alpha|^2 A_v^2}{\overline{V_{RS}^2}|\alpha|^2 A_v^2 + \overline{V_n^2}}. \tag{2.112}$$

It follows that

$$\text{NF} = \frac{V_{in}^2}{4kTR_S} \cdot \frac{\overline{V_{RS}^2}|\alpha|^2 A_v^2 + \overline{V_n^2}}{V_{in}^2 |\alpha|^2 A_v^2} \tag{2.113}$$

$$= \frac{1}{\overline{V_{RS}^2}} \cdot \frac{\overline{V_{RS}^2}|\alpha|^2 A_v^2 + \overline{V_n^2}}{|\alpha|^2 A_v^2} \tag{2.114}$$

$$= 1 + \frac{\overline{V_n^2}}{|\alpha|^2 A_v^2} \cdot \frac{1}{\overline{V_{RS}^2}}. \tag{2.115}$$

This result leads to another definition of the NF: the total noise at the output divided by the noise at the output due to the source impedance. The NF is usually specified for a 1-Hz bandwidth at a given frequency, and hence sometimes called the "spot noise figure" to emphasize the small bandwidth.

Equation (2.115) suggests that the NF depends on the *source impedance*, not only through $\overline{V_{RS}^2}$ but also through $\overline{V_n^2}$ (Example 2.19). In fact, if we model the noise by *input-referred* sources, then the input noise current, $\overline{I_{n,in}^2}$, partially flows through R_S, generating a source-dependent noise voltage of $\overline{I_{n,in}^2} R_S^2$ at the input and hence a proportional noise at the output. Thus, the NF must be specified with respect to a source impedance—typically 50 Ω.

For hand analysis and simulations, it is possible to reduce the right-hand side of Eq. (2.114) to a simpler form by noting that the numerator is the *total* noise measured at the output:

$$\text{NF} = \frac{1}{4kTR_S} \cdot \frac{\overline{V_{n,out}^2}}{A_0^2}, \tag{2.116}$$

where $\overline{V_{n,out}^2}$ includes both the source impedance noise and the LNA noise, and $A_0 = |\alpha| A_v$ is the voltage gain from V_{in} to V_{out} (rather than the gain from the LNA input to its output). We loosely say, "to calculate the NF, we simply divide the total output noise by the gain

from V_{in} to V_{out} and normalize the result to the noise of R_S." Alternatively, we can say from (2.115) that "we calculate the output noise due to the amplifier ($\overline{V_n^2}$), divide it by the gain, normalize it to $4kTR_S$, and add 1 to the result."

It is important to note that the above derivations are valid even if no actual *power* is transferred from the antenna to the LNA or from the LNA to a load. For example, if Z_{in} in Fig. 2.48(b) goes to infinity, no power is delivered to the LNA, but all of the derivations remain valid because they are based on *voltage* (squared) quantities rather than power quantities. In other words, so long as the derivations incorporate noise and signal voltages, no inconsistency arises in the presence of impedance mismatches or even infinite input impedances. This is a critical difference in thinking between modern RF design and traditional microwave design.

Example 2.20

Compute the noise figure of a shunt resistor R_P with respect to a source impedance R_S [Fig. 2.49(a)].

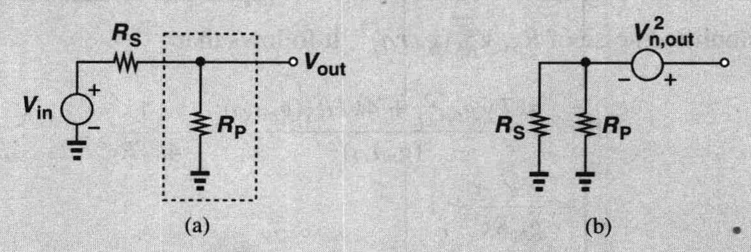


Figure 2.49 *(a) Circuit consisting of a single parallel resistor, (b) model for NF calculation.*

Solution:

From Fig. 2.49(b), the total output noise voltage is obtained by setting V_{in} to zero:

$$\overline{V_{n,out}^2} = 4kT(R_S||R_P). \tag{2.117}$$

The gain is equal to

$$A_0 = \frac{R_P}{R_P + R_S}. \tag{2.118}$$

Thus,

$$\text{NF} = 4kT(R_S||R_P)\frac{(R_S + R_P)^2}{R_P^2}\frac{1}{4kTR_S} \tag{2.119}$$

$$= 1 + \frac{R_S}{R_P}. \tag{2.120}$$

The NF is therefore minimized by *maximizing* R_P. Note that if $R_P = R_S$ to provide impedance matching, then the NF cannot be less than 3 dB. We will return to this critical point in the context of LNA design in Chapter 5.

Example 2.21

Determine the noise figure of the common-source stage shown in Fig. 2.50(a) with respect to a source impedance R_S. Neglect the capacitances and flicker noise of M_1 and assume I_1 is ideal.

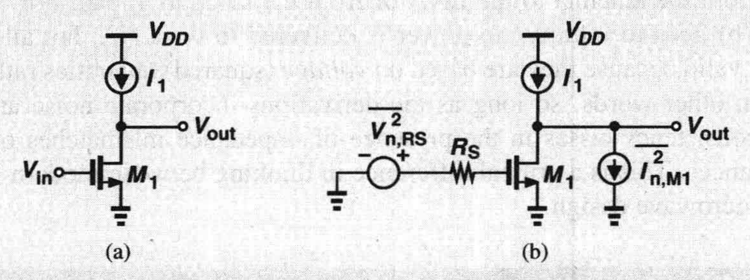


Figure 2.50 *(a) CS stage, (b) inclusion of noise.*

Solution:

From Fig. 2.50(b), the output noise consists of two components: (a) that due to M_1, $\overline{I^2_{n,M1}r_O^2}$, and (b) the amplified noise of R_S, $\overline{V^2_{RS}}(g_m r_O)^2$. It follows that

$$\text{NF} = \frac{4kT\gamma g_m r_O^2 + 4kTR_S(g_m r_O)^2}{(g_m r_O)^2} \cdot \frac{1}{4kTR_S} \tag{2.121}$$

$$= \frac{\gamma}{g_m R_S} + 1. \tag{2.122}$$

This result implies that the NF falls as R_S rises. Does this mean that, even though the amplifier remains unchanged, the overall system noise performance *improves* as R_S increases?! This interesting point is studied in Problem 2.16.

Noise Figure of Cascaded Stages Since many stages appear in a receiver chain, it is desirable to determine the NF of the overall cascade in terms of that of each stage. Consider the cascade depicted in Fig. 2.51(a), where A_{v1} and A_{v2} denote the *unloaded* voltage gain of the two stages. The input and output impedances and the output noise voltages of the two stages are also shown.[19]

We first obtain the NF of the cascade using a direct method; according to (2.115), we simply calculate the total noise at the output due to the two stages, divide by $(V_{out}/V_{in})^2$, normalize to $4kTR_S$, and add one to the result. Taking the loadings into account, we write the overall voltage gain as

$$A_0 = \frac{V_{out}}{V_{in}} = \frac{R_{in1}}{R_{in1} + R_S} A_{v1} \frac{R_{in2}}{R_{in2} + R_{out1}} A_{v2}. \tag{2.123}$$

19. We assume for simplicity that the reactive components of the input and output impedances are nulled but the final result is valid even if they are not.

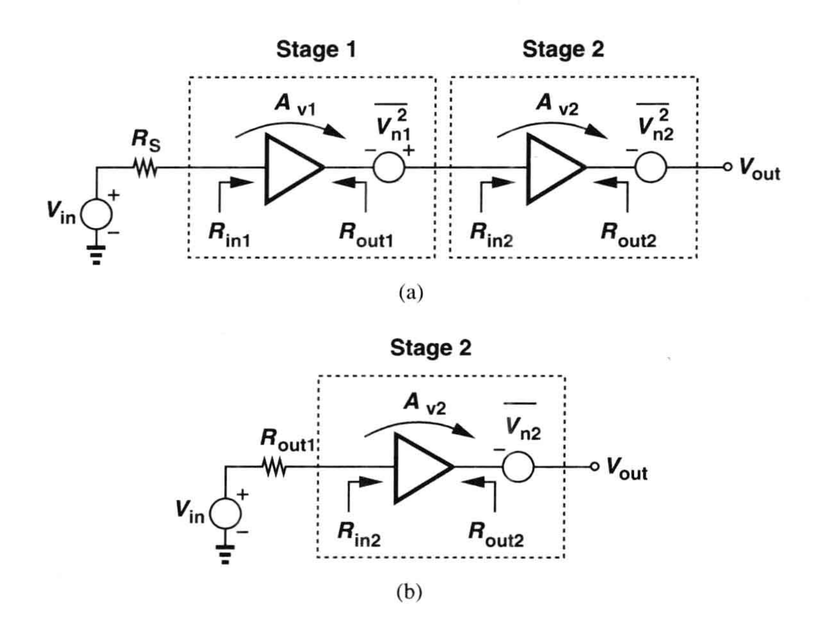


Figure 2.51 *(a) Noise in a cascade of stages, (b) simplified diagram.*

The output noise due to the two stages, denoted by $\overline{V_{n,out}^2}$, consists of two components: (a) $\overline{V_{n2}^2}$, and (b) $\overline{V_{n1}^2}$ amplified by the second stage. Since V_{n1} sees an impedance of R_{out1} to its left and R_{in2} to its right, it is scaled by a factor of $R_{in2}/(R_{in2} + R_{out1})$ as it appears at the input of the second stage. Thus,

$$\overline{V_{n,out}^2} = \overline{V_{n2}^2} + \overline{V_{n1}^2}\frac{R_{in2}^2}{(R_{in2} + R_{out1})^2}A_{v2}^2. \tag{2.124}$$

The overall NF is therefore expressed as

$$\mathrm{NF_{tot}} = 1 + \frac{\overline{V_{n,out}^2}}{A_0^2} \cdot \frac{1}{4kTR_S} \tag{2.125}$$

$$= 1 + \frac{\overline{V_{n1}^2}}{\left(\dfrac{R_{in1}}{R_{in1} + R_S}\right)^2 A_{v1}^2} \cdot \frac{1}{4kTR_S}$$

$$+ \frac{\overline{V_{n2}^2}}{\left(\dfrac{R_{in1}}{R_{in1} + R_S}\right)^2 A_{v1}^2 \left(\dfrac{R_{in2}}{R_{in2} + R_{out1}}\right)^2 A_{v2}^2} \cdot \frac{1}{4kTR_S} \tag{2.126}$$

The first two terms constitute the *NF* of the first stage, NF_1, with respect to a source impedance of R_S. The third term represents the noise of the second stage, but how can it be expressed in terms of the *noise figure* of this stage?

Let us now consider the second stage by itself and determine its noise figure with respect to a source impedance of R_{out1} [Fig. 2.51(b)]. Using (2.115) again, we have

$$NF_2 = 1 + \frac{\overline{V_{n2}^2}}{\frac{R_{in2}^2}{(R_{in2} + R_{out1})^2} A_{v2}^2} \frac{1}{4kTR_{out1}}. \tag{2.127}$$

It follows from (2.126) and (2.127) that

$$NF_{tot} = NF_1 + \frac{NF_2 - 1}{\frac{R_{in1}^2}{(R_{in1} + R_S)^2} A_{v1}^2 \frac{R_S}{R_{out1}}}. \tag{2.128}$$

What does the denominator represent? This quantity is in fact the "available power gain" of the first stage, defined as the "available power" at its output, $P_{out,av}$ (the power that it would deliver to a matched load) divided by the available source power, $P_{S,av}$ (the power that the source would deliver to a matched load). This can be readily verified by finding the power that the first stage in Fig. 2.51(a) would deliver to a load equal to R_{out1}:

$$P_{out,av} = V_{in}^2 \frac{R_{in1}^2}{(R_S + R_{in1})^2} A_{v1}^2 \cdot \frac{1}{4R_{out1}}. \tag{2.129}$$

Similarly, the power that V_{in} would deliver to a load of R_S is given by

$$P_{S,av} = \frac{V_{in}^2}{4R_S}. \tag{2.130}$$

The ratio of (2.129) and (2.130) is indeed equal to the denominator in (2.128).

With these observations, we write

$$NF_{tot} = NF_1 + \frac{NF_2 - 1}{A_{P1}}, \tag{2.131}$$

where A_{P1} denotes the "available power gain" of the first stage. It is important to bear in mind that NF_2 is computed with respect to the output impedance of the first stage. For m stages,

$$NF_{tot} = 1 + (NF_1 - 1) + \frac{NF_2 - 1}{A_{P1}} + \cdots + \frac{NF_m - 1}{A_{P1} \cdots A_{P(m-1)}}. \tag{2.132}$$

Called "Friis' equation" [7], this result suggests that the noise contributed by each stage decreases as the total gain preceding that stage increases, implying that the first few stages in a cascade are the most critical. Conversely, if a stage suffers from attenuation (loss), then the NF of the following circuits is "amplified" when referred to the input of that stage.

Example 2.22

Determine the *NF* of the cascade of common-source stages shown in Fig. 2.52. Neglect the transistor capacitances and flicker noise.

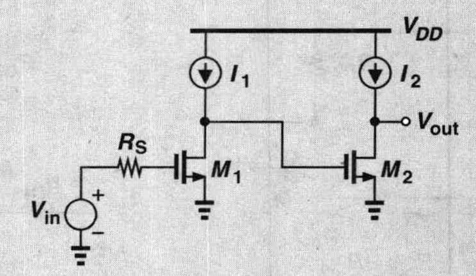


Figure 2.52 *Cascade of CS stages for noise figure calculation.*

Solution:

Which approach is simpler to use here, the direct method or Friis' equation? Since $R_{in1} = R_{in2} = \infty$, Eq. (2.126) reduces to

$$NF = 1 + \frac{\overline{V_{n1}^2}}{A_{v1}^2}\frac{1}{4kTR_S} + \frac{\overline{V_{n2}^2}}{A_{v1}^2 A_{v2}^2}\frac{1}{4kTR_S}, \tag{2.133}$$

where $\overline{V_{n1}^2} = 4kT\gamma g_{m1} r_{O1}^2$, $\overline{V_{n2}^2} = 4kT\gamma g_{m2} r_{O2}^2$, $A_{v1} = g_{m1} r_{O1}$, and $A_{v2} = g_{m2} r_{O2}$. With all of these quantities readily available, we simply substitute for their values in (2.133), obtaining

$$NF = 1 + \frac{\gamma}{g_{m1} R_S} + \frac{\gamma}{g_{m1}^2 r_{O1}^2 g_{m2} R_S}. \tag{2.134}$$

On the other hand, Friis' equation requires the calculation of the available power gain of the first stage and the *NF* of the second stage with respect to a source impedance of r_{O1}, leading to lengthy algebra.

The foregoing example represents a typical situation in modern RF design: the interface between the two stages does not have a 50-Ω impedance *and* no attempt has been made to provide impedance matching between the two stages. In such cases, Friis' equation becomes cumbersome, making direct calculation of the *NF* more attractive.

While the above example assumes an infinite input impedance for the second stage, the direct method can be extended to more realistic cases with the aid of Eq. (2.126). Even in the presence of complex input and output impedances, Eq. (2.126) indicates that (1) $\overline{V_{n1}^2}$ must be divided by the *unloaded* gain from V_{in} to the output of the first stage; (2) the output noise of the second stage, $\overline{V_{n2}^2}$, must be calculated with this stage driven by the output impedance of the first stage;[20] and (3) $\overline{V_{n2}^2}$ must be divided by the total voltage gain from V_{in} to V_{out}.

20. Recall from Example 2.19 that the output noise of a circuit may depend on the source impedance driving it, but the source impedance noise is excluded from $\overline{V_{n2}^2}$.

Example 2.23

Determine the noise figure of the circuit shown in Fig. 2.53(a). Neglect transistor capacitances, flicker noise, channel-length modulation, and body effect.

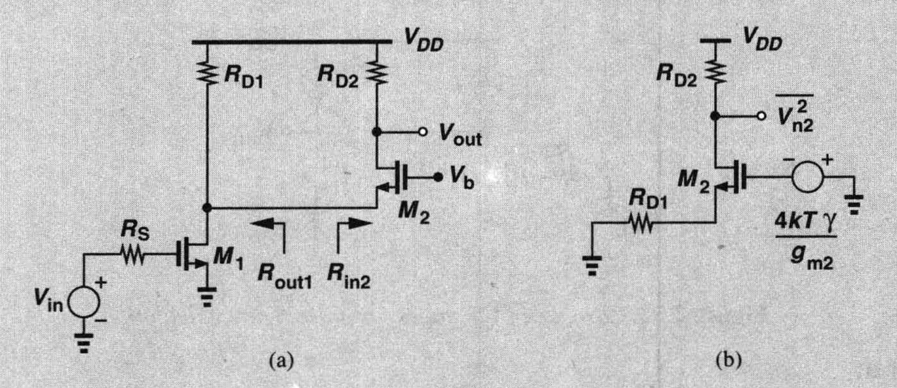


Figure 2.53 *(a) Cascade of CS and CG stages, (b) simplified diagram.*

Solution:

For the first stage, $A_{v1} = -g_{m1}R_{D1}$ and the unloaded output noise is equal to

$$\overline{V_{n1}^2} = 4kT\gamma g_{m1}R_{D1}^2 + 4kTR_{D1}. \tag{2.135}$$

For the second stage, the reader can show from Fig. 2.53(b) that

$$\overline{V_{n2}^2} = \frac{4kT\gamma}{g_{m2}}\left(\frac{R_{D2}}{\frac{1}{g_{m2}} + R_{D1}}\right)^2 + 4kTR_{D2}. \tag{2.136}$$

Note that the output impedance of the first stage is included in the calculation of $\overline{V_{n2}^2}$ but the *noise* of R_{D1} is not.

We now substitute these values in Eq. (2.126), bearing in mind that $R_{in2} = 1/g_{m2}$ and $A_{v2} = g_{m2}R_{D2}$.

$$NF_{tot} = 1 + \frac{4kT\gamma g_{m1}R_D^2 + 4kTR_{D1}}{g_{m1}^2 R_{D1}^2} \cdot \frac{1}{4kTR_S} + \frac{\frac{4kT\gamma}{g_{m2}}\left(\frac{R_{D2}}{g_{m2}^{-1} + R_{D2}}\right)^2 + 4kTR_{D2}}{g_{m1}^2 R_{D1}^2 \left(\frac{g_{m2}^{-1}}{g_{m2}^{-1} + R_{D1}}\right)^2 g_{m2}^2 R_{D2}^2} \cdot \frac{1}{4kTR_S}. \tag{2.137}$$

Noise Figure of Lossy Circuits Passive circuits such as filters appear at the front end of RF transceivers and their loss proves critical (Chapter 4). The loss arises from unwanted

resistive components within the circuit that convert the input power to heat, thereby producing a smaller signal power at the output. Furthermore, recall from Fig. 2.37 that resistive components also *generate* thermal noise. That is, passive lossy circuits both attenuate the signal and introduce noise.

We wish to prove that the noise figure of a passive (reciprocal) circuit is equal to its "power loss," defined as $L = P_{in}/P_{out}$, where P_{in} is the available source power and P_{out} the available power at the output. As mentioned in the derivation of Friis' equation, the available power is the power that a given source or circuit would deliver to a conjugate-matched load. The proof is straightforward if the input and output are matched (Problem 2.17). We consider a more general case here.

Consider the arrangement shown in Fig. 2.54(a), where the lossy circuit is driven by a source impedance of R_S while driving a load impedance of R_L.[21] From Eq. (2.130), the available source power is $P_{in} = V_{in}^2/(4R_S)$. To determine the available output power, we construct the Thevenin equivalent shown in Fig. 2.54(b), obtaining $P_{out} = V_{Thev}^2/(4R_{out})$. Thus, the loss is given by

$$L = \frac{V_{in}^2}{V_{Thev}^2}\frac{R_{out}}{R_S}. \tag{2.138}$$

To calculate the noise figure, we utilize the theorem illustrated in Fig. 2.37 and the equivalent circuit shown in Fig. 2.54(c) to write

$$\overline{V_{n,out}^2} = 4kTR_{out}\frac{R_L^2}{(R_L + R_{out})^2}. \tag{2.139}$$

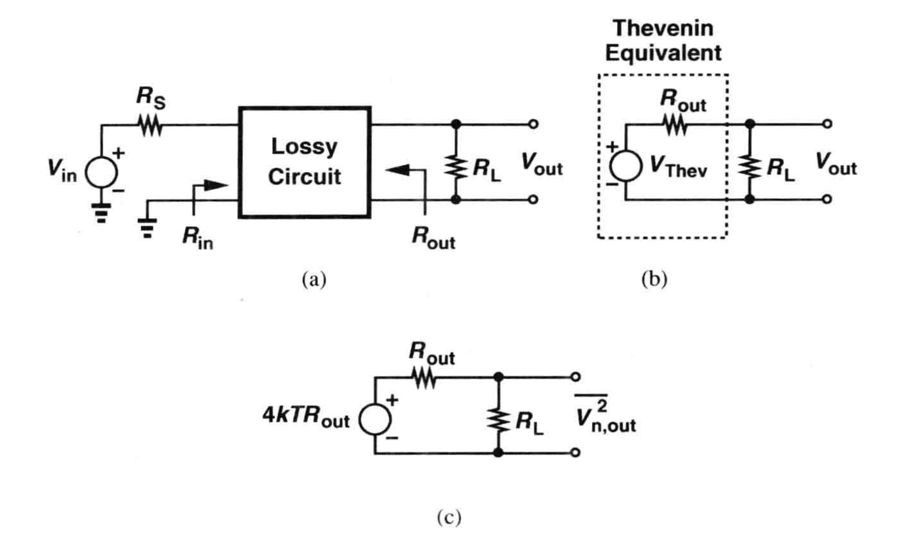


Figure 2.54 *(a) Lossy passive network, (b) Thevenin equivalent, (c) simplified diagram.*

21. For simplicity, we assume the reactive parts of the impedances are cancelled but the final result is valid even if they are not.

Note that R_L is assumed noiseless so that only the noise figure of the lossy circuit can be determined. The voltage gain from V_{in} to V_{out} is found by noting that, in response to V_{in}, the circuit produces an output voltage of $V_{out} = V_{Thev}R_L/(R_L + R_{out})$ [Fig. 2.54(b)]. That is,

$$A_0 = \frac{V_{Thev}}{V_{in}} \frac{R_L}{R_L + R_{out}}. \quad (2.140)$$

The NF is equal to (2.139) divided by the square of (2.140) and normalized to $4kTR_S$:

$$NF = 4kTR_{out} \frac{V_{in}^2}{V_{Thev}^2} \frac{1}{4kTR_S} \quad (2.141)$$

$$= L. \quad (2.142)$$

Example 2.24

The receiver shown in Fig. 2.55 incorporates a front-end band-pass filter (BPF) to suppress some of the interferers that may desensitize the LNA. If the filter has a loss of L and the LNA a noise figure of NF_{LNA}, calculate the overall noise figure.

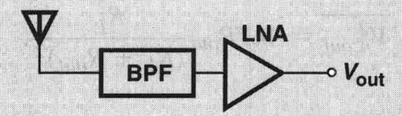


Figure 2.55 *Cascade of BPF and LNA.*

Solution:

Denoting the noise figure of the filter by NF_{filt}, we write Friis' equation as

$$NF_{tot} = NF_{filt} + \frac{NF_{LNA} - 1}{L^{-1}} \quad (2.143)$$

$$= L + (NF_{LNA} - 1)L \quad (2.144)$$

$$= L \cdot NF_{LNA}, \quad (2.145)$$

where NF_{LNA} is calculated with respect to the output resistance of the filter. For example, if $L = 1.5$ dB and $NF_{LNA} = 2$ dB, then $NF_{tot} = 3.5$ dB.

2.4 SENSITIVITY AND DYNAMIC RANGE

The performance of RF receivers is characterized by many parameters. We study two, namely, sensitivity and dynamic range, here and defer the others to Chapter 3.

2.4.1 Sensitivity

The sensitivity is defined as the minimum signal level that a receiver can detect with "acceptable quality." In the presence of excessive noise, the detected signal becomes unintelligible and carries little information. We define acceptable quality as sufficient signal-to-noise ratio, which itself depends on the type of modulation and the corruption (e.g., bit error rate) that the system can tolerate. Typical required SNR levels are in the range of 6 to 25 dB (Chapter 3).

In order to calculate the sensitivity, we write

$$NF = \frac{SNR_{in}}{SNR_{out}} \tag{2.146}$$

$$= \frac{P_{sig}/P_{RS}}{SNR_{out}}, \tag{2.147}$$

where P_{sig} denotes the input signal power and P_{RS} the source resistance noise power, both per unit bandwidth. Do we express these quantities in V^2/Hz or W/Hz? Since the input impedance of the receiver is typically matched to that of the antenna (Chapter 4), the antenna indeed delivers signal power and noise power to the receiver. For this reason, it is common to express both quantities in W/Hz (or dBm/Hz). It follows that

$$P_{sig} = P_{RS} \cdot NF \cdot SNR_{out}. \tag{2.148}$$

Since the overall signal power is distributed across a certain bandwidth, B, the two sides of (2.148) must be integrated over the bandwidth so as to obtain the total mean squared power. Assuming a flat spectrum for the signal and the noise, we have

$$P_{sig,tot} = P_{RS} \cdot NF \cdot SNR_{out} \cdot B. \tag{2.149}$$

Equation (2.149) expresses the sensitivity as the minimum input signal that yields a given value for the output SNR. Changing the notation slightly and expressing the quantities in dB or dBm, we have[22]

$$P_{sen}|_{dBm} = P_{RS}|_{dBm/Hz} + NF|_{dB} + SNR_{min}|_{dB} + 10 \log B, \tag{2.150}$$

where P_{sen} is the sensitivity and B is expressed in Hz. Note that (2.150) does not directly depend on the gain of the system. If the receiver is matched to the antenna, then from (2.91), $P_{RS} = kT = -174$ dBm/Hz and

$$P_{sen} = -174\,\text{dBm/Hz} + NF + 10 \log B + SNR_{min}. \tag{2.151}$$

Note that the sum of the first three terms is the total integrated noise of the system (sometimes called the "noise floor").

22. Note that in conversion to dB or dBm, we take 10 log because these are power quantities.

Example 2.25

A GSM receiver requires a minimum *SNR* of 12 dB and has a channel bandwidth of 200 kHz. A wireless LAN receiver, on the other hand, specifies a minimum *SNR* of 23 dB and has a channel bandwidth of 20 MHz. Compare the sensitivities of these two systems if both have an *NF* of 7 dB.

Solution:

For the GSM receiver, $P_{sen} = -102$ dBm, whereas for the wireless LAN system, $P_{sen} = -71$ dBm. Does this mean that the latter is inferior? No, the latter employs a much wider bandwidth and a more efficient modulation to accommodate a data rate of 54 Mb/s. The GSM system handles a data rate of only 270 kb/s. In other words, specifying the sensitivity of a receiver without the data rate is not meaningful.

2.4.2 Dynamic Range

Dynamic range (DR) is loosely defined as the maximum input level that a receiver can "tolerate" divided by the minimum input level that it can detect (the sensitivity). This definition is quantified differently in different applications. For example, in analog circuits such as analog-to-digital converters, the DR is defined as the "full-scale" input level divided by the input level at which $SNR = 1$. The full scale is typically the input level beyond which a hard saturation occurs and can be easily determined by examining the circuit.

In RF design, on the other hand, the situation is more complicated. Consider a simple common-source stage. How do we define the input "full scale" for such a circuit? Is there a particular input level beyond which the circuit becomes excessively nonlinear? We may view the 1-dB compression point as such a level. But, what if the circuit senses two interferers and suffers from intermodulation?

In RF design, two definitions of DR have emerged. The first, simply called the dynamic range, refers to the maximum tolerable *desired* signal power divided by the minimum tolerable desired signal power (the sensitivity). Illustrated in Fig. 2.56(a), this DR is limited by compression at the upper end and noise at the lower end. For example, a cell phone coming close to a base station may receive a very large signal and must process it with

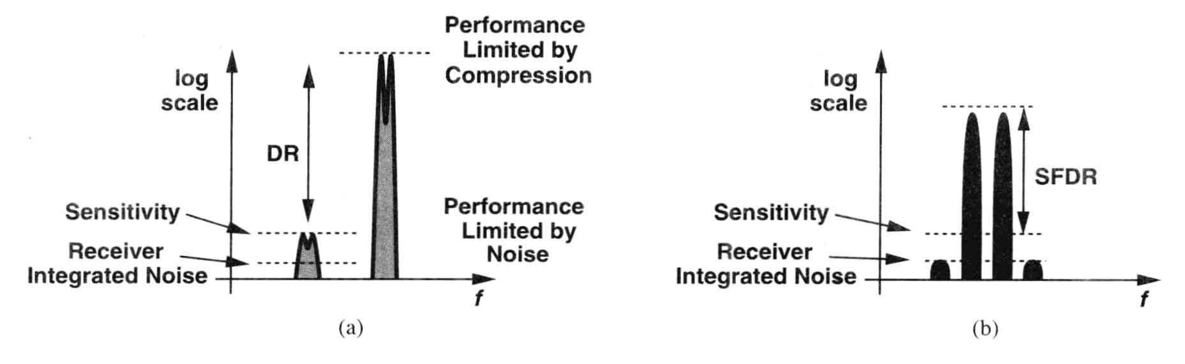


Figure 2.56 *Definitions of (a) DR and (b) SFDR.*

acceptable distortion. In fact, the cell phone measures the signal strength and adjusts the receiver gain so as to avoid compression. Excluding interferers, this "compression-based" DR can exceed 100 dB because the upper end can be raised relatively easily.

The second type, called the "spurious-free dynamic range" (SFDR), represents limitations arising from both noise and interference. The lower end is still equal to the sensitivity, but the upper end is defined as the maximum input level in a *two-tone* test for which the third-order IM products do not exceed the integrated noise of the receiver. As shown in Fig. 2.56(b), two (modulated or unmodulated) tones having equal amplitudes are applied and their level is raised until the IM products reach the integrated noise.[23] The ratio of the power of each tone to the sensitivity yields the SFDR. The SFDR represents the maximum relative level of interferers that a receiver can tolerate while producing an acceptable signal quality from a small input level.

Where should the various levels depicted in Fig. 2.56(b) be measured, at the input of the circuit or at its output? Since the IM components appear only at the output, the output port serves as a more natural candidate for such a measurement. In this case, the sensitivity—usually an input-referred quantity—must be scaled by the gain of the circuit so that it is referred to the output. Alternatively, the output IM magnitudes can be divided by the gain so that they are referred to the input. We follow the latter approach in our SFDR calculations.

To determine the upper end of the SFDR, we rewrite Eq. (2.56) as

$$P_{IIP3} = P_{in} + \frac{P_{out} - P_{IM,out}}{2}, \tag{2.152}$$

where, for the sake of brevity, we have denoted $20 \log A_x$ as P_x even though no actual power may be transferred at the input or output ports. Also, $P_{IM,out}$ represents the level of IM products at the output. If the circuit exhibits a gain of G (in dB), then we can refer the IM level to the input by writing $P_{IM,in} = P_{IM,out} - G$. Similarly, the *input* level of each tone is given by $P_{in} = P_{out} - G$. Thus, (2.152) reduces to

$$P_{IIP3} = P_{in} + \frac{P_{in} - P_{IM,in}}{2} \tag{2.153}$$

$$= \frac{3P_{in} - P_{IM,in}}{2}, \tag{2.154}$$

and hence

$$P_{in} = \frac{2P_{IIP3} + P_{IM,in}}{3}. \tag{2.155}$$

The upper end of the SFDR is that value of P_{in} which makes $P_{IM,in}$ equal to the integrated noise of the receiver:

$$P_{in,max} = \frac{2P_{IIP3} + (-174\,\text{dBm} + NF + 10 \log B)}{3}. \tag{2.156}$$

23. Note that the integrated noise is a single value (e.g., 100 μV_{rms}), not a *density*.

The SFDR is the difference (in dB) between $P_{in,max}$ and the sensitivity:

$$SFDR = P_{in,max} - (-174\,\text{dBm} + NF + 10\log B + SNR_{min}) \tag{2.157}$$

$$= \frac{2(P_{IIP3} + 174\,\text{dBm} - NF - 10\log B)}{3} - SNR_{min}. \tag{2.158}$$

For example, a GSM receiver with $NF = 7\,\text{dB}$, $P_{IIP3} = -15\,\text{dBm}$, and $SNR_{min} = 12\,\text{dB}$ achieves an SFDR of 54 dB, a substantially lower value than the dynamic range in the absence of interferers.

Example 2.26

The upper end of the dynamic range is limited by intermodulation in the presence of *two* interferers or desensitization in the presence of *one* interferer. Compare these two cases and determine which one is more restrictive.

Solution:

We must compare the upper end expressed by Eq. (2.156) with the 1-dB compression point:

$$P_{1-dB} \overset{?}{\gtrless} P_{in,max}. \tag{2.159}$$

Since $P_{1-dB} = P_{IIP3} - 9.6\,\text{dB}$,

$$P_{IIP3} - 9.6\,\text{dB} \overset{?}{\gtrless} \frac{2P_{IIP3} + (-174\,\text{dBm} + NF + 10\log B)}{3} \tag{2.160}$$

and hence

$$P_{IIP3} - 28.8\,\text{dB} \overset{?}{\gtrless} -174\,\text{dBm} + NF + 10\log B. \tag{2.161}$$

Since the right-hand side represents the receiver noise floor, we expect it to be much lower than the left-hand side. In fact, even for an extremely wideband channel of $B = 1$ GHz and $NF = 10\,\text{dB}$, the right-hand side is equal to $-74\,\text{dBm}$, whereas, with a typical P_{IIP3} of -10 to $-25\,\text{dBm}$, the left-hand side still remains higher. It is therefore plausible to conclude that

$$P_{1-dB} > P_{in,max}. \tag{2.162}$$

It follows that the maximum tolerable level in a two-tone test is quite lower than that in a compression test, i.e., corruption by intermodulation between two interferers is much greater than compression due to one. The SFDR is therefore a more stringent characteristic of the system than the compression-based dynamic range.

2.5 PASSIVE IMPEDANCE TRANSFORMATION

At radio frequencies, we often employ passive networks to transform impedances—from high to low and vice versa, or from complex to real and vice versa. Called "matching

networks," such circuits do not easily lend themselves to integration because their constituent devices, particularly inductors, suffer from loss if built on silicon chips. (We do use on-chip inductors in many RF building blocks.) Nonetheless, a basic understanding of impedance transformation is essential.

2.5.1 Quality Factor

In its simplest form, the quality factor, Q, indicates how close to ideal an energy-storing device is. An ideal capacitor dissipates no energy, exhibiting an infinite Q, but a series resistance, R_S [Fig. 2.57(a)], reduces its Q to

$$Q_S = \frac{\frac{1}{C\omega}}{R_S}, \tag{2.163}$$

where the numerator denotes the "desired" component and the denominator, the "undesired" component. If the resistive loss in the capacitor is modeled by a *parallel* resistance [Fig. 2.57(b)], then we must define the Q as

$$Q_P = \frac{R_P}{\frac{1}{C\omega}}, \tag{2.164}$$

because an ideal (infinite Q) results only if $R_P = \infty$. As depicted in Figs. 2.57(c) and (d), similar concepts apply to inductors

$$Q_S = \frac{L\omega}{R_S} \tag{2.165}$$

$$Q_P = \frac{R_P}{L\omega}. \tag{2.166}$$

While a parallel resistance appears to have no physical meaning, modeling the loss by R_P proves useful in many circuits such as amplifiers and oscillators (Chapters 5 and 8). We will also introduce other definitions of Q in Chapter 8.

2.5.2 Series-to-Parallel Conversion

Before studying transformation techniques, let us consider the series and parallel *RC* sections shown in Fig. 2.58. What choice of values makes the two networks equivalent?

R_S C (a) R_P C (b) R_S L (c) R_P L (d)

Figure 2.57 *(a) Series RC circuit, (b) equivalent parallel circuit, (c) series RL circuit, (d) equivalent parallel circuit.*

Figure 2.58 *Series-to-parallel conversion.*

Equating the impedances,

$$\frac{R_S C_S s + 1}{C_S s} = \frac{R_P}{R_P C_P s + 1}, \tag{2.167}$$

and substituting $j\omega$ for s, we have

$$R_P C_S j\omega = 1 - R_P C_P R_S C_S \omega^2 + (R_P C_P + R_S C_S) j\omega, \tag{2.168}$$

and hence

$$R_P C_P R_S C_S \omega^2 = 1 \tag{2.169}$$

$$R_P C_P + R_S C_S - R_P C_S = 0. \tag{2.170}$$

Equation (2.169) implies that $Q_S = Q_P$.

Of course, the two impedances cannot remain equal at all frequencies. For example, the series section approaches an open circuit at low frequencies while the parallel section does not. Nevertheless, an approximation allows equivalence for a narrow frequency range. We first substitute for $R_P C_P$ in (2.169) from (2.170), obtaining

$$R_P = \frac{1}{R_S C_S^2 \omega^2} + R_S. \tag{2.171}$$

Utilizing the definition of Q_S in (2.163), we have

$$R_P = (Q_S^2 + 1) R_S. \tag{2.172}$$

Substitution in (2.169) thus yields

$$C_P = \frac{Q_S^2}{Q_S^2 + 1} C_S. \tag{2.173}$$

So long as $Q_S^2 \gg 1$ (which is true for a finite frequency range),

$$R_P \approx Q_S^2 R_S \tag{2.174}$$

$$C_P \approx C_S. \tag{2.175}$$

That is, the series-to-parallel conversion retains the value of the capacitor but raises the resistance by a factor of Q_S^2. These approximations for R_P and C_P are relatively accurate because the quality factors encountered in practice typically exceed 4. Conversely,

parallel-to-series conversion reduces the resistance by a factor of Q_P^2. This statement applies to *RL* sections as well.

2.5.3 Basic Matching Networks

A common situation in RF transmitter design is that a load resistance must be transformed to a lower value. The circuit shown in Fig. 2.59(a) accomplishes this task. As mentioned above, the capacitor in parallel with R_L converts this resistance to a lower *series* component [Fig. 2.59(b)]. The inductance is inserted to cancel the equivalent series capacitance.

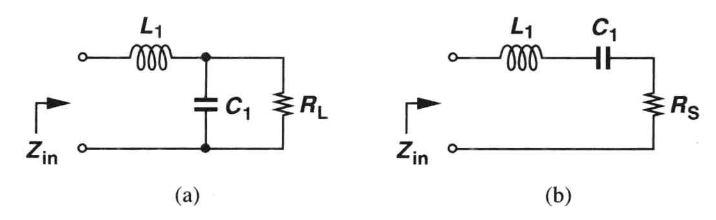


Figure 2.59 *(a) Matching network, (b) equivalent circuit.*

Writing Z_{in} from Fig. 2.59(a) and replacing s with $j\omega$, we have

$$Z_{in}(j\omega) = \frac{R_L(1 - L_1C_1\omega^2) + jL_1\omega}{1 + jR_LC_1\omega}. \tag{2.176}$$

Thus,

$$Re\{Z_{in}\} = \frac{R_L}{1 + R_L^2C_1^2\omega^2} \tag{2.177}$$

$$= \frac{R_L}{1 + Q_P^2}, \tag{2.178}$$

indicating that R_L is transformed down by a factor of $1 + Q_P^2$. Also, setting the imaginary part to zero gives

$$L_1 = \frac{R_L^2C_1}{1 + R_L^2C_1^2\omega^2} \tag{2.179}$$

$$= \frac{R_L^2C_1}{1 + Q_P^2}. \tag{2.180}$$

If $Q_P^2 \gg 1$, then

$$Re\{Z_{in}\} \approx \frac{1}{R_LC_1^2\omega^2} \tag{2.181}$$

$$L_1 = \frac{1}{C_1\omega^2}. \tag{2.182}$$

The following example illustrates how the component values are chosen.

Example 2.27

Design the matching network of Fig. 2.59(a) so as to transform $R_L = 50\ \Omega$ to $25\ \Omega$ at a center frequency of 5 GHz.

Solution:

Assuming $Q_P^2 \gg 1$, we have from Eqs. (2.181) and (2.182), $C_1 = 0.90$ pF and $L_1 = 1.13$ nH, respectively. Unfortunately, however, $Q_P = 1.41$, indicating that Eqs. (2.178) and (2.180) must be used instead. We thus obtain $C_1 = 0.637$ pF and $L_1 = 0.796$ nH.

In order to transform a resistance to a higher value, the capacitive network shown in Fig. 2.60(a) can be used. The series-parallel conversion results derived previously provide insight here. If $Q^2 \gg 1$, the parallel combination of C_1 and R_L can be converted to a series network [Fig. 2.60(b)], where $R_S \approx [R_L(C_1\omega)^2]^{-1}$ and $C_S \approx C_1$. Viewing C_2 and C_1 as one capacitor, C_{eq}, and converting the resulting series section to a parallel circuit [Fig. 2.60(c)], we have

$$R_{tot} = \frac{1}{R_S(C_{eq}\omega)^2} \tag{2.183}$$

$$= \left(1 + \frac{C_1}{C_2}\right)^2 R_L. \tag{2.184}$$

That is, the network "boosts" the value of R_L by a factor of $(1 + C_1/C_2)^2$. Also,

$$C_{eq} = \frac{C_1 C_2}{C_1 + C_2}. \tag{2.185}$$

Note that the capacitive component must be cancelled by placing an inductor in *parallel* with the input.

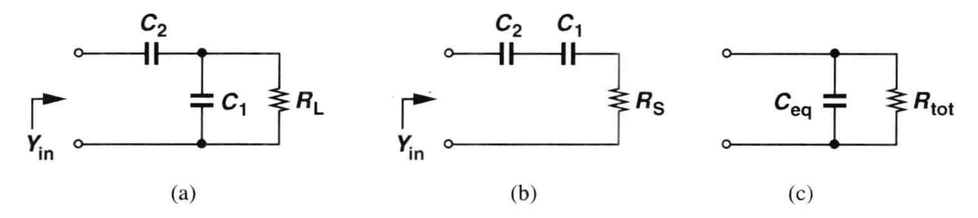


Figure 2.60 *(a) Capacitive matching circuit, (b) simplified circuit with parallel-to-series conversion, (c) simplified circuit with series-to-parallel conversion.*

For low Q values, the above derivations incur significant error. We thus compute the input admittance $(1/Y_{in})$ and replace s with $j\omega$,

$$Y_{in} = \frac{j\omega C_2(1 + j\omega R_L C_1)}{1 + R_L(C_1 + C_2)j\omega}. \tag{2.186}$$

The real part of Y_{in} yields the equivalent resistance seen to ground if we write

$$R_{tot} = \frac{1}{Re\{Y_{in}\}} \tag{2.187}$$

$$= \frac{1}{R_L C_2^2 \omega^2} + R_L \left(1 + \frac{C_1}{C_2}\right)^2. \tag{2.188}$$

In comparison with Eq. (2.184), this result contains an additional component, $(R_L C_2^2 \omega^2)^{-1}$.

Example 2.28

Determine how the circuit shown in Fig. 2.61(a) transforms R_L.

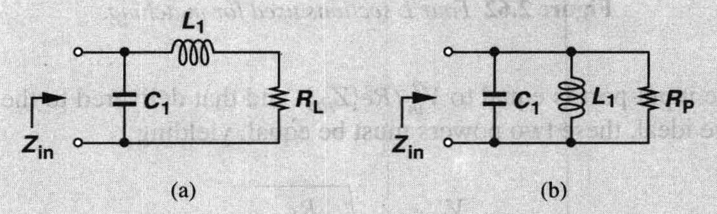


Figure 2.61 *(a) Matching network, (b) simplified circuit.*

Solution:

We postulate that conversion of the L_1–R_L branch to a parallel section produces a higher resistance. If $Q_S^2 = (L_1\omega/R_L)^2 \gg 1$, then the equivalent parallel resistance is obtained from Eq. (2.174) as

$$R_P = Q_S^2 R_L \tag{2.189}$$

$$= \frac{L_1^2 \omega^2}{R_L}. \tag{2.190}$$

The parallel equivalent inductance is approximately equal to L_1 and is cancelled by C_1 [Fig. 2.61(b)].

The intuition gained from our analysis of matching networks leads to the four "L-section" topologies[24] shown in Fig. 2.62. In Fig. 2.62(a), C_1 transforms R_L to a smaller series value and L_1 cancels C_1. Similarly, in Fig. 2.62(b), L_1 transforms R_L to a smaller series value while C_1 resonates with L_1. In Fig. 2.62(c), L_1 transforms R_L to a larger parallel value and C_1 cancels the resulting parallel inductance. A similar observation applies to Fig. 2.62(d).

How do these networks transform voltages and currents? As an example, consider the circuit in Fig. 2.62(a). For a sinusoidal input voltage with an rms value of V_{in}, the power

24. The term "L" is used because the capacitor and the inductor form the letter L in the circuit diagram.

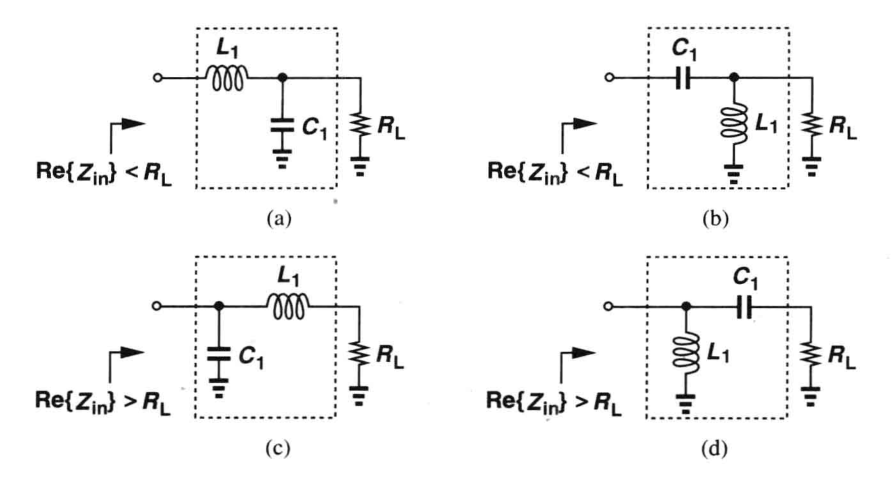


Figure 2.62 *Four L sections used for matching.*

delivered to the input port is equal to $V_{in}^2/Re\{Z_{in}\}$, and that delivered to the load, V_{out}^2/R_L. If L_1 and C_1 are ideal, these two powers must be equal, yielding

$$\frac{V_{out}}{V_{in}} = \sqrt{\frac{R_L}{Re\{Z_{in}\}}}. \tag{2.191}$$

This result, of course, applies to any lossless matching network whose input impedance contains a zero imaginary part. Since $P_{in} = V_{in}I_{in}$ and $P_{out} = V_{out}I_{out}$, we also have

$$\frac{I_{out}}{I_{in}} = \sqrt{\frac{Re\{Z_{in}\}}{R_L}}. \tag{2.192}$$

For example, a network transforming R_L to a *lower* value "amplifies" the voltage and attenuates the current by the above factor.

Example 2.29

A closer look at the L-sections in Figs. 2.62(a) and (c) suggests that one can be obtained from the other by swapping the input and output ports. Is it possible to generalize this observation?

Solution:

Yes, it is. Consider the arrangement shown in Fig. 2.63(a), where the passive network transforms R_L by a factor of α. Assuming the input port exhibits no imaginary component, we equate the power delivered to the network to the power delivered to the load:

$$\left(V_{in}\frac{\alpha R_L}{\alpha R_L + R_S}\right)^2 \cdot \frac{1}{\alpha R_L} = \frac{V_{out}^2}{R_L}. \tag{2.193}$$

Example 2.29 (*Continued*)

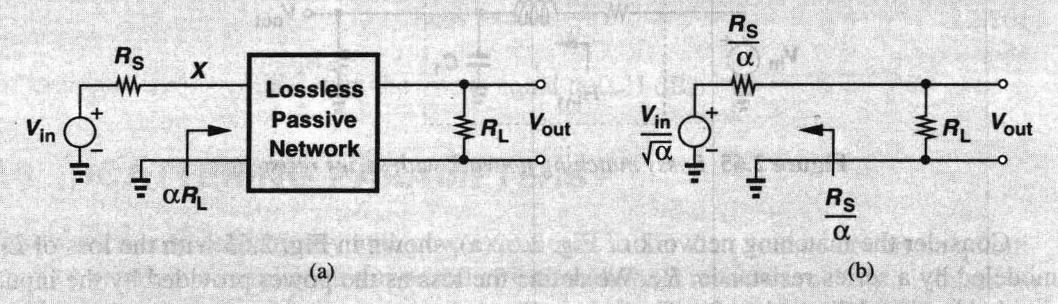


Figure 2.63 *(a) Input and (b) output impedances of a lossless passive network.*

It follows that

$$V_{out} = \frac{V_{in}}{\sqrt{\alpha}} \cdot \frac{R_L}{R_L + \frac{R_S}{\alpha}}, \tag{2.194}$$

pointing to the Thevenin equivalent shown in Fig. 2.63(b). We observe that the network transforms R_S by a factor of $1/\alpha$ and the input voltage by a factor of $1/\sqrt{\alpha}$, similar to that in Eq. (2.191). In other words, if the input and output ports of such a network are swapped, the resistance transformation ratio is simply inverted.

Transformers can also transform impedances. An ideal transformer having a turns ratio of n "amplifies" the input voltage by a factor of n (Fig. 2.64). Since no power is lost, $V_{in}^2/R_{in} = n^2 V_{in}^2/R_L$ and hence $R_{in} = R_L/n^2$. The behavior of actual transformers, especially those fabricated monolithically, is studied in Chapter 7.

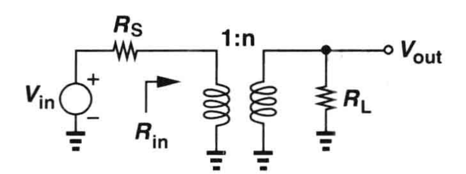


Figure 2.64 *Impedance transformation by a physical transformer.*

The networks studied here operate across only a narrow bandwidth because the transformation ratio, e.g., $1 + Q^2$, varies with frequency, and the capacitance and inductance approximately resonate over a narrow frequency range. Broadband matching networks can be constructed, but they typically suffer from a high loss.

2.5.4 Loss in Matching Networks

Our study of matching networks has thus far neglected the loss of their constituent components, particularly, that of inductors. We analyze the effect of loss in a few cases here, but, in general, simulations are necessary to determine the behavior of complex lossy networks.

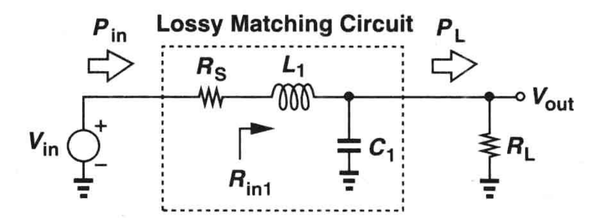


Figure 2.65 *Lossy matching network with series resistence.*

Consider the matching network of Fig. 2.62(a), shown in Fig. 2.65 with the loss of L_1 modeled by a series resistance, R_S. We define the loss as the power provided by the input divided by that delivered to R_L. The former is equal to

$$P_{in} = \frac{V_{in}^2}{R_S + R_{in1}} \tag{2.195}$$

and the latter,

$$P_L = \left(V_{in}\frac{R_{in1}}{R_S + R_{in1}}\right)^2 \cdot \frac{1}{R_{in1}}, \tag{2.196}$$

because the power delivered to R_{in1} is entirely absorbed by R_L. It follows that

$$\text{Loss} = \frac{P_{in}}{P_L} \tag{2.197}$$

$$= 1 + \frac{R_S}{R_{in1}}. \tag{2.198}$$

For example, if $R_S = 0.1R_{in1}$, then the (power) loss reaches 0.41 dB. Note that this network transforms R_L to a *lower* value, $R_{in1} = R_L/(1 + Q_P^2)$, thereby suffering from loss even if R_S appears small.

As another example, consider the network of Fig. 2.62(b), depicted in Fig. 2.66 with the loss of L_1 modeled by a parallel resistance, R_P. We note that the power delivered by V_{in}, P_{in}, is entirely absorbed by $R_P||R_L$:

$$P_{in} = \frac{V_{out}^2}{R_P||R_L} \tag{2.199}$$

$$= \frac{V_{out}^2}{R_L}\frac{R_P + R_L}{R_P}. \tag{2.200}$$

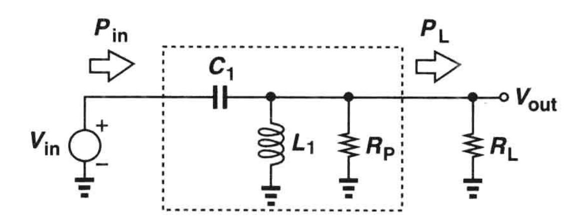


Figure 2.66 *Lossy matching network with parallel resistence.*

Recognizing V_{out}^2/R_L as the power delivered to the load, P_L, we have

$$\text{Loss} = 1 + \frac{R_L}{R_P}. \tag{2.201}$$

For example, if $R_P = 10R_L$, then the loss is equal to 0.41 dB.

2.6 SCATTERING PARAMETERS

Microwave theory deals mostly with power quantities rather than voltage or current quantities. Two reasons can explain this approach. First, traditional microwave design is based on transfer of *power* from one stage to the next. Second, the measurement of high-frequency voltages and currents in the laboratory proves very difficult, whereas that of average power is more straightforward. Microwave theory therefore models devices, circuits, and systems by parameters that can be obtained through the measurement of power quantities. They are called "scattering parameters" (S-parameters).

Before studying S-parameters, we introduce an example that provides a useful viewpoint. Consider the L_1–C_1 series combination depicted in Fig. 2.67. The circuit is driven by a sinusoidal source, V_{in}, having an output impedance of R_S. A load resistance of $R_L = R_S$ is tied to the output port. At an input frequency of $\omega = (\sqrt{L_1C_1})^{-1}$, L_1 and C_1 form a short circuit, providing a conjugate match between the source and the load. In analogy with transmission lines, we say the "incident wave" produced by the signal source is absorbed by R_L. At other frequencies, however, L_1 and C_1 attenuate the voltage delivered to R_L. Equivalently, we say the input port of the circuit generates a "reflected wave" that returns to the source. In other words, the difference between the incident power (the power that would be delivered to a matched load) and the reflected power represents the power delivered to the circuit.

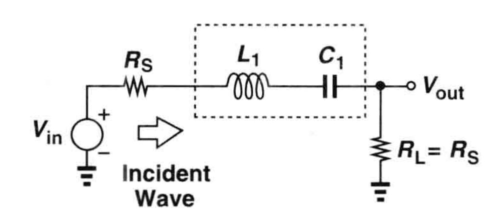


Figure 2.67 *Incident wave in a network.*

The above viewpoint can be generalized for any two-port network. As illustrated in Fig. 2.68, we denote the incident and reflected waves at the input port by V_1^+ and V_1^-, respectively. Similar waves are denoted by V_2^+ and V_2^-, respectively, at the output. Note

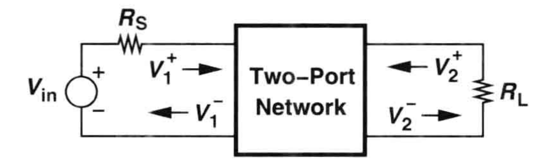


Figure 2.68 *Illustration of incident and reflected waves at the input and output.*

that V_1^+ denotes a wave generated by V_{in} as if the input impedance of the circuit were equal to R_S. Since that may not be the case, we include the reflected wave, V_1^-, so that the actual voltage measured at the input is equal to $V_1^+ + V_1^-$. Also, V_2^+ denotes the incident wave traveling *into* the output port or, equivalently, the wave *reflected* from R_L. These four quantities are uniquely related to one another through the S-parameters of the network:

$$V_1^- = S_{11}V_1^+ + S_{12}V_2^+ \quad (2.202)$$

$$V_2^- = S_{21}V_1^+ + S_{22}V_2^+. \quad (2.203)$$

With the aid of Fig. 2.69, we offer an intuitive interpretation for each parameter:

1. For S_{11}, we have from Fig. 2.69(a)

$$S_{11} = \frac{V_1^-}{V_1^+}|_{V_2^+=0}. \quad (2.204)$$

   Thus, S_{11} is the ratio of the reflected and incident waves at the input port when the reflection from R_L (i.e., V_2^+) is zero. This parameter represents the accuracy of the input matching.

2. For S_{12}, we have from Fig. 2.69(b)

$$S_{12} = \frac{V_1^-}{V_2^+}|_{V_1^+=0}. \quad (2.205)$$

   Thus, S_{12} is the ratio of the reflected wave at the input port to the incident wave into the output port when the input port is matched. In this case, the *output* port is driven by the signal source. This parameter characterizes the "reverse isolation" of the circuit, i.e., how much of the output signal couples to the input network.

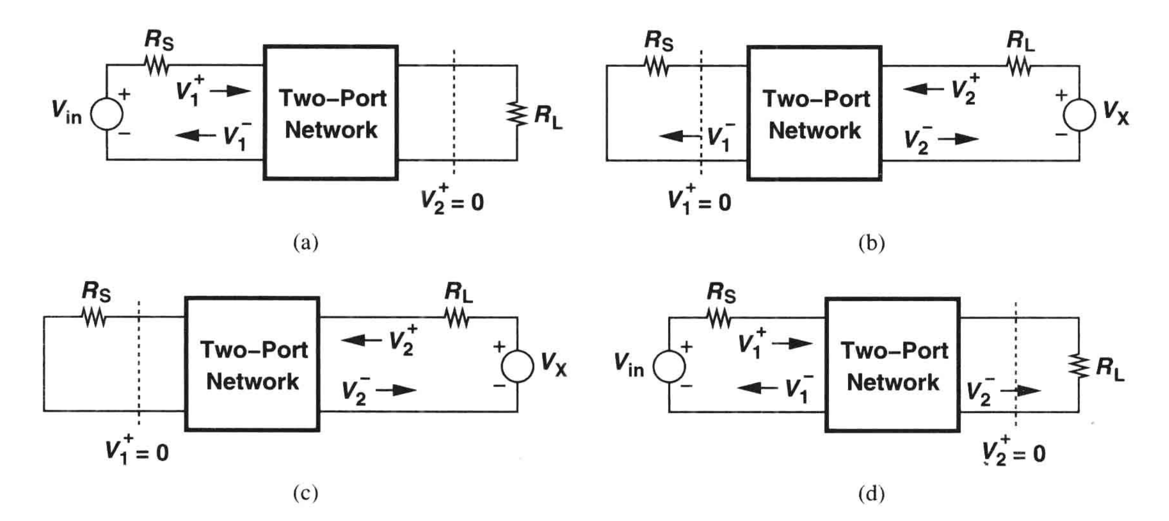


Figure 2.69 *Illustration of four S-parameters.*

3. For S_{22}, we have from Fig. 2.69(c)

$$S_{22} = \frac{V_2^-}{V_2^+}|_{V_1^+=0}. \tag{2.206}$$

Thus, S_{22} is the ratio of reflected and incident waves at the output when the reflection from R_S (i.e., V_1^+) is zero. This parameter represents the accuracy of the output matching.

4. For S_{21}, we have from Fig. 2.69(d)

$$S_{21} = \frac{V_2^-}{V_1^+}|_{V_2^+=0}. \tag{2.207}$$

Thus, S_{21} is the ratio of the wave incident on the load to that going to the input when the reflection from R_L is zero. This parameter represents the gain of the circuit.

We should make a few remarks at this point. First, S-parameters generally have frequency-dependent complex values. Second, we often express S-parameters in units of dB as follows

$$S_{mn}|_{dB} = 20\log|S_{mn}|. \tag{2.208}$$

Third, the condition $V_2^+ = 0$ in Eqs. (2.204) and (2.207) requires that the reflection from R_L be zero, but it does *not* mean that the output port of the circuit must be conjugate-matched to R_L. This condition simply means that if, hypothetically, a transmission line having a characteristic impedance equal to R_S carries the output signal to R_L, then no wave is reflected from R_L. A similar note applies to the requirement $V_1^+ = 0$ in Eqs. (2.205) and (2.206). The conditions $V_1^+ = 0$ at the input or $V_2^+ = 0$ at the output facilitate high-frequency measurements while creating issues in modern RF design. As mentioned in Section 2.3.5 and exemplified by the cascade of stages in Fig. 2.53, modern RF design typically does not strive for matching between stages. Thus, if S_{11} of the first stage must be measured with $R_L = R_S$ at its output, then its value may not represent the S_{11} of the cascade.

In modern RF design, S_{11} is the most commonly-used S parameter as it quantifies the accuracy of impedance matching at the input of receivers. Consider the arrangement shown in Fig. 2.70, where the receiver exhibits an input impedance of Z_{in}. The incident wave V_1^+ is given by $V_{in}/2$ (as if Z_{in} were equal to R_S). Moreover, the total voltage at the receiver

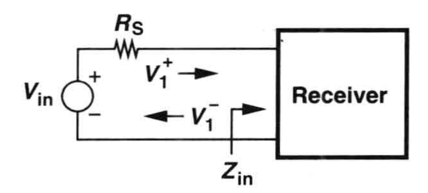


Figure 2.70 *Receiver with incident and reflected waves.*

input is equal to $V_{in}Z_{in}/(Z_{in} + R_S)$, which is also equal to $V_1^+ + V_1^-$. Thus,

$$V_1^- = V_{in}\frac{Z_{in}}{Z_{in} + R_S} - \frac{V_{in}}{2} \tag{2.209}$$

$$= \frac{Z_{in} - R_S}{2(Z_{in} + R_S)}V_{in}. \tag{2.210}$$

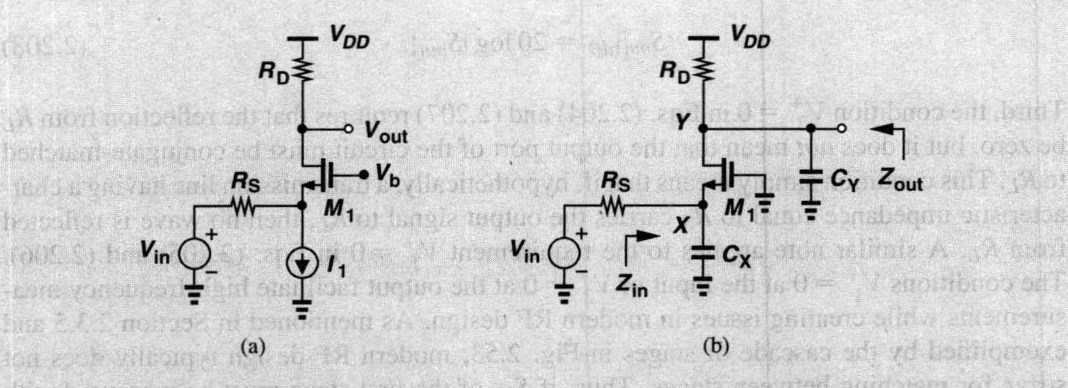
It follows that

$$\frac{V_1^-}{V_1^+} = \frac{Z_{in} - R_S}{Z_{in} + R_S}. \tag{2.211}$$

Called the "input reflection coefficient" and denoted by Γ_{in}, this quantity can also be considered to be S_{11} if we remove the condition $V_2^+ = 0$ in Eq. (2.204).

Example 2.30

Determine the S-parameters of the common-gate stage shown in Fig. 2.71(a). Neglect channel-length modulation and body effect.



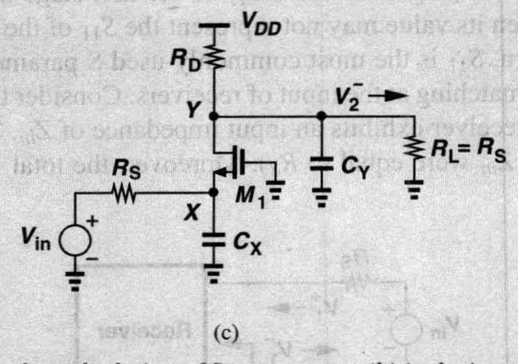


Figure 2.71 *(a) CG stage for calculation of S-parameters, (b) inclusion of capacitors, (c) effect of reflected wave at output.*

Example 2.30 *(Continued)*

Solution:

Drawing the circuit as shown in Fig. 2.71(b), where $C_X = C_{GS} + C_{SB}$ and $C_Y = C_{GD} + C_{DB}$, we write $Z_{in} = (1/g_m)||(C_X s)^{-1}$ and

$$S_{11} = \frac{Z_{in} - R_S}{Z_{in} + R_S} \tag{2.212}$$

$$= \frac{1 - g_m R_S - C_X s}{1 + g_m R_S + C_X s}. \tag{2.213}$$

For S_{12}, we recognize that the arrangement of Fig. 2.71(b) yields no coupling from the output to the input if channel-length modulation is neglected. Thus, $S_{12} = 0$. For S_{22}, we note that $Z_{out} = R_D||(C_Y s)^{-1}$ and hence

$$S_{22} = \frac{Z_{out} - R_S}{Z_{out} + R_S} \tag{2.214}$$

$$= -\frac{R_S - R_D + R_S R_D C_Y s}{R_S + R_D + R_S R_D C_Y s}. \tag{2.215}$$

Lastly, S_{21} is obtained according to the configuration of Fig. 2.71(c). Since $V_2^-/V_{in} = (V_2^-/V_X)(V_X/V_{in})$, $V_2^-/V_X = g_m[R_D||R_S||(C_Y s)^{-1}]$, and $V_X/V_{in} = Z_{in}/(Z_{in} + R_S)$, we obtain

$$\frac{V_2^-}{V_{in}} = g_m\left(R_D||R_S||\frac{1}{C_Y s}\right)\frac{1}{1 + g_m R_S + R_S C_X s}. \tag{2.216}$$

It follows that

$$S_{21} = 2g_m\left(R_D||R_S||\frac{1}{C_Y s}\right)\frac{1}{1 + g_m R_S + R_S C_X s}. \tag{2.217}$$

2.7 ANALYSIS OF NONLINEAR DYNAMIC SYSTEMS[25]

In our treatment of systems in Section 2.2, we have assumed a static nonlinearity, e.g., in the form of $y(t) = \alpha_1 x(t) + \alpha_2 x^2(t) + \alpha_3 x^3(t)$. In some cases, a circuit may exhibit dynamic nonlinearity, requiring a more complex analysis. In this section, we address this task.

2.7.1 Basic Considerations

Let us first consider a general nonlinear system with an input given by $x(t) = A_1 \cos\omega_1 t + A_2 \cos\omega_2 t$. We expect the output, $y(t)$, to contain harmonics at $n\omega_1$, $m\omega_2$, and IM products

25. This section can be skipped in a first reading.

at $k\omega_1 \pm q\omega_2$, where, n, m, k, and q are integers. In other words,

$$y(t) = \sum_{n=1}^{\infty} a_n \cos(n\omega_1 t + \theta_n) + \sum_{n=1}^{\infty} b_n \cos(n\omega_2 t + \phi_n) + \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} c_{m,n} \cos(n\omega_1 t + m\omega_2 t + \phi_{n,m}). \tag{2.218}$$

In the above equation, a_n, b_n, $c_{m,n}$, and the phase shifts are frequency-dependent quantities. If the differential equation governing the system is known, we can simply substitute for $y(t)$ from this expression, equate the like terms, and compute a_n, b_n, $c_{m,n}$, and the phase shifts. For example, consider the simple RC section shown in Fig. 2.72, where the capacitor is nonlinear and expressed as $C_1 = C_0(1 + \alpha V_{out})$. Adding the voltages across R_1 and C_1 and equating the result to V_{in}, we have

$$R_1 C_0(1 + \alpha V_{out})\frac{dV_{out}}{dt} + V_{out} = V_{in}. \tag{2.219}$$

Now suppose $V_{in}(t) = V_0 \cos \omega_1 t + V_0 \cos \omega_2 t$ (as in a two-tone test) and assume the system is only "weakly" nonlinear, i.e., only the output terms at ω_1, ω_2, $\omega_1 \pm \omega_2$, $2\omega_1 \pm \omega_2$, and $2\omega_2 \pm \omega_1$ are significant. Thus, the output assumes the form

$$\begin{aligned} V_{out}(t) = {} & a_1 \cos(\omega_1 t + \phi_1) + b_1 \cos(\omega_2 t + \phi_2) + c_1 \cos[(\omega_1 + \omega_2)t + \phi_3] \\ & + c_2 \cos[(\omega_1 - \omega_2)t + \phi_4] + c_3 \cos[(2\omega_1 + \omega_2)t + \phi_5] \\ & + c_4 \cos[(\omega_1 + 2\omega_2)t + \phi_6] + c_5 \cos[(2\omega_1 - \omega_2)t + \phi_7] \\ & + c_6 \cos[(\omega_1 - 2\omega_2)t + \phi_8], \end{aligned} \tag{2.220}$$

where, for simplicity, we have used c_m and ϕ_m. We must now substitute for $V_{out}(t)$ and $V_{in}(t)$ in (2.219), convert products of sinusoids to sums, bring all of the terms to one side of the equation, group them according to their frequencies, and equate the coefficient of each sinusoid to zero. We thus obtain a system of 16 nonlinear equations and 16 knowns (a_1, b_1, $c_1, \ldots, c_6, \phi_1, \ldots, \phi_8$).

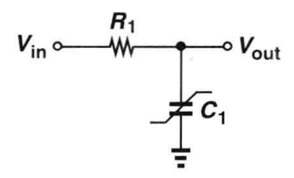


Figure 2.72 *RC circuit with nonlinear capacitor.*

This type of analysis is called "harmonic balance" because it predicts the output frequencies and attempts to "balance" the two sides of the circuit's differential equation by including these components in $V_{out}(t)$. The mathematical labor in harmonic balance makes hand analysis difficult or impossible. The "Volterra series" approach, on the other hand, prescribes a *recursive* method that computes the response more accurately in successive

steps *without* the need for solving nonlinear equations. A detailed treatment of the concepts described below can be found in [10–14].

2.8 VOLTERRA SERIES

In order to understand how the Volterra series represents the time response of a system, we begin with a simple input form, $V_{in}(t) = V_0 \exp(j\omega_1 t)$. Of course, if we wish to obtain the response to a sinusoid of the form $V_0 \cos \omega_1 t = Re\{V_0 \exp(j\omega_1 t)\}$, we simply calculate the real part of the output.[26] (The use of the exponential form greatly simplifies the manipulation of the product terms.) For a linear, time-invariant system, the output is given by

$$V_{out}(t) = H(\omega_1) V_0 \exp(j\omega_1 t), \tag{2.221}$$

where $H(\omega_1)$ is the Fourier transform of the impulse response. For example, if the capacitor in Fig. 2.72 is linear, i.e., $C_1 = C_0$, then we can substitute for V_{out} and V_{in} in Eq. (2.219):

$$R_1 C_0 H(\omega_1)(j\omega_1) V_0 \exp(j\omega_1 t) + H(\omega_1) V_0 \exp(j\omega_1 t) = V_0 \exp(j\omega_1 t). \tag{2.222}$$

It follows that

$$H(\omega_1) = \frac{1}{R_1 C_0 j\omega_1 + 1}. \tag{2.223}$$

Note that the phase shift introduced by the circuit is included in $H(\omega_1)$ here.

As our next step, let us ask, how should the output response of a dynamic nonlinear system be expressed? To this end, we apply two tones to the input, $V_{in}(t) = V_0 \exp(j\omega_1 t) + V_0 \exp(j\omega_2 t)$, recognizing that the output consists of both linear and nonlinear responses. The former are of the form

$$V_{out1}(t) = H(\omega_1) V_0 \exp(j\omega_1 t) + H(\omega_2) V_0 \exp(j\omega_2 t), \tag{2.224}$$

and the latter include exponentials such as $\exp[j(\omega_1 + \omega_2)t]$, etc. We expect that the coefficient of such an exponential is a function of both ω_1 and ω_2. We thus make a slight change in our notation: we denote $H(\omega_j)$ in Eq. (2.224) by $H_1(\omega_j)$ [to indicate first-order (linear) terms] and the coefficient of $\exp[j(\omega_1 + \omega_2)t]$ by $H_2(\omega_1, \omega_2)$. In other words, the overall output can be written as

$$\begin{aligned} V_{out}(t) &= H_1(\omega_1) V_0 \exp(j\omega_1 t) + H_1(\omega_2) V_0 \exp(j\omega_2 t) \\ &\quad + H_2(\omega_1, \omega_2) V_0^2 \exp[j(\omega_1 + \omega_2)t] + \cdots . \end{aligned} \tag{2.225}$$

How do we determine the terms at $2\omega_1$, $2\omega_2$, and $\omega_1 - \omega_2$? If $H_2(\omega_1, \omega_2)\exp[j(\omega_1 + \omega_2)t]$ represents the component at $\omega_1 + \omega_2$, then $H_2(\omega_1, \omega_1)\exp[j(2\omega_1)t]$ must model

26. From another point of view, in $V_0 \exp(j\omega_1 t) = V_0 \cos \omega_1 t + jV_0 \sin \omega_1 t$, the first term generates its own response, as does the second term; the two responses remain distinguishable by virtue of the factor j.

that at $2\omega_1$. Similarly, $H_2(\omega_2, \omega_2)$ and $H_2(\omega_1, -\omega_2)$ serve as coefficients for $\exp[j(2\omega_2)t]$ and $\exp[j(\omega_1 - \omega_2)t]$, respectively. In other words, a more complete form of Eq. (2.225) reads

$$\begin{aligned} V_{out}(t) = & H_1(\omega_1)V_0 \exp(j\omega_1 t) + H_1(\omega_2)V_0 \exp(j\omega_2 t) + H_2(\omega_1, \omega_1)V_0^2 \exp(2j\omega_1 t) \\ & + H_2(\omega_2, \omega_2)V_0^2 \exp(2j\omega_2 t) + H_2(\omega_1, \omega_2)V_0^2 \exp[j(\omega_1 + \omega_2)t] \\ & + H_2(\omega_1, -\omega_2)V_0^2 \exp[j(\omega_1 - \omega_2)t] + \cdots . \end{aligned} \quad (2.226)$$

Thus, our task is simply to compute $H_2(\omega_1, \omega_2)$.

Example 2.31

Determine $H_2(\omega_1, \omega_2)$ for the circuit of Fig. 2.72.

Solution:

We apply the input $V_{in}(t) = V_0 \exp(j\omega_1 t) + V_0 \exp(j\omega_2 t)$ and assume the output is of the form $V_{out}(t) = H_1(\omega_1)V_0 \exp(j\omega_1 t) + H_1(\omega_2)V_0 \exp(j\omega_2 t) + H_2(\omega_1, \omega_2)V_0^2 \exp[j(\omega_1 + \omega_2)t]$. We substitute for V_{out} and V_{in} in Eq. (2.219):

$$\begin{aligned} & R_1C_0[1 + \alpha H_1(\omega_1)V_0 e^{j\omega_1 t} + \alpha H_1(\omega_2)V_0 e^{j\omega_2 t} + \alpha H_2(\omega_1, \omega_2)V_0^2 e^{j(\omega_1+\omega_2)t}] \\ & \quad \times [H_1(\omega_1)j\omega_1 V_0 e^{j\omega_1 t} + H_1(\omega_2)j\omega_2 V_0 e^{j\omega_2 t} + H_2(\omega_1, \omega_2)j(\omega_1 + \omega_2) \\ & \quad \times V_0^2 e^{j(\omega_1+\omega_2)t}] + H_1(\omega_1)e^{j\omega_1 t} + H_1(\omega_2)e^{j\omega_2 t} + H_2(\omega_1, \omega_2)V_0^2 e^{j(\omega_1+\omega_2)t} \\ & \quad = V_0 e^{j\omega_1 t} + V_0 e^{j\omega_2 t}. \end{aligned} \quad (2.227)$$

To obtain H_2, we only consider the terms containing $\omega_1 + \omega_2$:

$$\begin{aligned} & R_1C_0[\alpha H_1(\omega_1)H_1(\omega_2)j\omega_1 V_0^2 e^{j(\omega_1+\omega_2)t} + \alpha H_1(\omega_2)H_1(\omega_1)j\omega_2 V_0^2 e^{j(\omega_1+\omega_2)t} \\ & \quad + H_2(\omega_1, \omega_2)j(\omega_1 + \omega_2)V_0^2 e^{j(\omega_1+\omega_2)t}] + H_2(\omega_1, \omega_2) \\ & \quad \times V_0^2 e^{j(\omega_1+\omega_2)t} = 0 \end{aligned} \quad (2.228)$$

That is,

$$H_2(\omega_1, \omega_2) = -\frac{\alpha R_1 C_0 j(\omega_1 + \omega_2)H_1(\omega_1)H_1(\omega_2)}{R_1 C_0 j(\omega_1 + \omega_2) + 1}. \quad (2.229)$$

Noting that the denominator resembles that of (2.223) but with ω_1 replaced by $\omega_1 + \omega_2$, we simplify $H_2(\omega_1, \omega_2)$ to

$$H_2(\omega_1, \omega_2) = -\alpha R_1 C_0 j(\omega_1 + \omega_2)H_1(\omega_1)H_1(\omega_2)H_1(\omega_1 + \omega_2). \quad (2.230)$$

Why did we assume $V_{out}(t) = H_1(\omega_1)V_0 \exp(j\omega_1 t) + H_1(\omega_2)V_0 \exp(j\omega_2 t) + H_2 V_0^2(\omega_1, \omega_2) \exp[j(\omega_1 + \omega_2)t]$ while we know that $V_{out}(t)$ also contains terms at $2\omega_1$, $2\omega_2$, and $\omega_1 - \omega_2$? This is because these other exponentials do not yield terms of the form $\exp[j(\omega_1 + \omega_2)t]$.

Example 2.32

If an input $V_0 \exp(j\omega_1 t)$ is applied to the circuit of Fig. 2.72, determine the amplitude of the second harmonic at the output.

Solution:

As mentioned earlier, the component at $2\omega_1$ is obtained as $H_2(\omega_1, \omega_1)V_0^2 \exp[j(\omega_1 + \omega_1)t]$. Thus, the amplitude is equal to

$$|A_{2\omega 1}| = |\alpha R_1 C_0 (2\omega_1) H_1^2(\omega_1) H_1(2\omega_1)| V_0^2 \tag{2.231}$$

$$= \frac{2|\alpha| R_1 C_0 \omega_1 V_0^2}{(R_1^2 C_0^2 \omega_1^2 + 1)\sqrt{4R_1^2 C_0^2 \omega_1^2 + 1}}. \tag{2.232}$$

We observe that $A_{2\omega 1}$ falls to zero as ω_1 approaches zero because C_1 draws little current, and also as ω_1 goes to infinity because the second harmonic is suppressed by the low-pass nature of the circuit.

Example 2.33

If two tones of equal amplitude are applied to the circuit of Fig. 2.72, determine the ratio of the amplitudes of the components at $\omega_1 + \omega_2$ and $\omega_1 - \omega_2$. Recall that $H_1(\omega) = (R_1 C_0 j\omega + 1)^{-1}$.

Solution:

From Eq. (2.230), the ratio is given by

$$\left|\frac{A_{\omega 1+\omega 2}}{A_{\omega 1-\omega 2}}\right| = \left|\frac{H_2(\omega_1, \omega_2)}{H_2(\omega_1, -\omega_2)}\right| \tag{2.233}$$

$$= \left|\frac{(\omega_1 + \omega_2)H_1(\omega_2)H_1(\omega_1 + \omega_2)}{(\omega_1 - \omega_2)H_1(-\omega_2)H_1(\omega_1 - \omega_2)}\right|. \tag{2.234}$$

Since $|H_1(\omega_2)| = |H_1(-\omega_2)|$, we have

$$\left|\frac{A_{\omega 1+\omega 2}}{A_{\omega 1-\omega 2}}\right| = \frac{(\omega_1 + \omega_2)\sqrt{R_1^2 C_0^2 (\omega_1 - \omega_2)^2 + 1}}{|\omega_1 - \omega_2|\sqrt{R_1^2 C_0^2 (\omega_1 + \omega_2)^2 + 1}}. \tag{2.235}$$

The foregoing examples point to a methodical approach that allows us to compute the second harmonic or second-order IM components with a moderate amount of algebra. But how about higher-order harmonics or IM products? We surmise that for Nth-order terms, we must apply the input $V_{in}(t) = V_0 \exp(j\omega_1 t) + \cdots + V_0 \exp(j\omega_N t)$ and compute $H_n(\omega_1, \ldots, \omega_n)$ as the coefficient of the $\exp[j(\omega_1 + \cdots + \omega_n)t]$ terms in the output. The

output can therefore be expressed as

$$V_{out}(t) = \sum_{k=1}^{N} H_1(\omega_k)V_0 \exp(j\omega_k t) + \sum_{m=1}^{N}\sum_{k=1}^{N} H_2(\omega_m, \pm\omega_k)V_0^2 \exp[j(\omega_m \pm \omega_k)t]$$
$$+ \sum_{n=1}^{N}\sum_{m=1}^{N}\sum_{k=1}^{N} H_3(\omega_n, \pm\omega_m, \pm\omega_k)V_0^3 \exp[j(\omega_n \pm \omega_m \pm \omega_k)t] + \cdots. \quad (2.236)$$

The above representation of the output is called the Volterra series. As exemplified by (2.230), $H_m(\omega_1, \ldots, \omega_m)$ can be computed in terms of $H_1, \ldots, H_{m-1}$ with no need to solve nonlinear equations. We call H_m the m-th "Volterra kernel."

Example 2.34

Determine the third Volterra kernel for the circuit of Fig. 2.72.

Solution:

We assume $V_{in}(t) = V_0 \exp(j\omega_1 t) + V_0 \exp(j\omega_2 t) + V_0 \exp(j\omega_3 t)$. Since the output contains many components, we introduce the short hands $H_{1(1)} = H_1(\omega_1)V_0 \exp(j\omega_1 t)$, $H_{1(2)} = H_1(\omega_2)V_0 \exp(j\omega_2 t)$, etc., $H_{2(1,2)} = H_2(\omega_1, \omega_2)V_0^2 \exp[j(\omega_1 + \omega_2)t]$, etc., and $H_{3(1,2,3)} = H_3(\omega_1, \omega_2, \omega_3)V_0^3 \exp[j(\omega_1 + \omega_2 + \omega_3)t]$. We express the output as

$$V_{out}(t) = H_{1(1)} + H_{1(2)} + H_{1(3)} + H_{2(1,2)} + H_{2(1,3)} + H_{2(2,3)} + H_{2(1,1)}$$
$$+ H_{2(2,2)} + H_{2(3,3)} + H_{3(1,2,3)} + \cdots. \quad (2.237)$$

We must substitute for V_{out} and V_{in} in Eq. (2.219) and group all of the terms that contain $\omega_1 + \omega_2 + \omega_3$. To obtain such terms in the product of αV_{out} and dV_{out}/dt, we note that $\alpha H_{2(1,2)} j\omega_3 H_{1(3)}$ and $\alpha H_{1(3)} j(\omega_1 + \omega_2) H_{2(1,2)}$ produce an exponential of the form $\exp[j(\omega_1 + \omega_2)t]\exp(j\omega_3)$. Similarly, $\alpha H_{2(2,3)} j\omega_1 H_{1(1)}$, $\alpha H_{1(1)} j(\omega_2 + \omega_3) H_{2(2,3)}$, $\alpha H_{2(1,3)} j\omega_2 H_{1(2)}$, and $\alpha H_{1(2)} j(\omega_1 + \omega_3) H_{2(1,3)}$ result in $\omega_1 + \omega_2 + \omega_3$. Finally, the product of αV_{out} and dV_{out}/dt also contains $1 \times j(\omega_1 + \omega_2 + \omega_3)H_{3(1,2,3)}$. Grouping all of the terms, we have

$$H_3(\omega_1, \omega_2, \omega_3)$$
$$= -j\alpha R_1 C_0 \frac{H_2(\omega_1, \omega_2)\omega_3 H_1(\omega_3) + H_2(\omega_2, \omega_3)\omega_1 H_1(\omega_1) + H_2(\omega_1, \omega_3)\omega_2 H_1(\omega_2)}{R_1 C_0 j(\omega_1 + \omega_2 + \omega_3) + 1}$$
$$- j\alpha R_1 C_0 \frac{H_1(\omega_1)(\omega_2 + \omega_3)H_2(\omega_2, \omega_3) + H_1(\omega_2)(\omega_1 + \omega_3)H_2(\omega_1, \omega_3)}{R_1 C_0 j(\omega_1 + \omega_2 + \omega_3) + 1}$$
$$- j\alpha R_1 C_0 \frac{H_1(\omega_3)(\omega_1 + \omega_2)H_2(\omega_1, \omega_2)}{R_1 C_0 j(\omega_1 + \omega_2 + \omega_3) + 1}. \quad (2.238)$$

Note that $H_{2(1,1)}$, etc., do not appear here and could have been omitted from Eq. (2.237). With the third Volterra kernel available, we can compute the amplitude of critical terms. For example, the third-order IM components in a two-tone test are obtained by substituting ω_1 for ω_3 and $-\omega_2$ for ω_2.

The reader may wonder if the Volterra series can be used with inputs other than exponentials. This is indeed possible [14] but beyond the scope of this book.

The approach described in this section is called the "harmonic" method of kernel calculation. In summary, this method proceeds as follows:

1. Assume $V_{in}(t) = V_0 \exp(j\omega_1 t)$ and $V_{out}(t) = H_1(\omega_1)V_0 \exp(j\omega_1 t)$. Substitute for V_{out} and V_{in} in the system's differential equation, group the terms that contain $\exp(j\omega_1 t)$, and compute the first (linear) kernel, $H_1(\omega_1)$.
2. Assume $V_{in}(t) = V_0 \exp(j\omega_1 t) + V_0 \exp(j\omega_2 t)$ and $V_{out}(t) = H_1(\omega_1)V_0 \exp(j\omega_1 t) + H_1(\omega_2)V_0 \exp(j\omega_2 t) + H_2(\omega_1, \omega_2)V_0^2 \exp[j(\omega_1 + \omega_2)t]$. Make substitutions in the differential equation, group the terms that contain $\exp[j(\omega_1 + \omega_2)t]$, and determine the second kernel, $H_2(\omega_1, \omega_2)$.
3. Assume $V_{in}(t) = V_0 \exp(j\omega_1 t) + V_0 \exp(j\omega_2 t) + V_0 \exp(j\omega_3 t)$ and $V_{out}(t)$ is given by Eq. (2.237). Make substitutions, group the terms that contain $\exp[j(\omega_1 + \omega_2 + \omega_3)t]$, and calculate the third kernel, $H_3(\omega_1, \omega_2, \omega_3)$.
4. To compute the amplitude of harmonics and IM components, choose $\omega_1, \omega_2, \ldots$ properly. For example, $H_2(\omega_1, \omega_1)$ yields the transfer function for $2\omega_1$ and $H_3(\omega_1, -\omega_2, \omega_1)$ the transfer function for $2\omega_1 - \omega_2$.

2.8.1 Method of Nonlinear Currents

As seen in Example 2.34, the harmonic method becomes rapidly more complex as n increases. An alternative approach called the method of "nonlinear currents" is sometimes preferred as it reduces the algebra to some extent. We describe the method itself here and refer the reader to [13] for a formal proof of its validity.

The method of nonlinear currents proceeds as follows for a circuit that contains a two-terminal nonlinear device [13]:

1. Assume $V_{in}(t) = V_0 \exp(j\omega_1 t)$ and determine the linear response of the circuit by ignoring the nonlinearity. The "response" includes both the output of interest *and* the voltage across the nonlinear device.
2. Assume $V_{in}(t) = V_0 \exp(j\omega_1 t) + V_0 \exp(j\omega_2 t)$ and calculate the voltage across the nonlinear device, assuming it is linear. Now, compute the *nonlinear* component of the current flowing through the device, assuming the device is nonlinear.
3. Set the main input to *zero* and place a current source equal to the nonlinear component found in Step 2 in parallel with the nonlinear device.
4. Ignoring the nonlinearity of the device again, determine the circuit's response to the current source applied in Step 3. Again, the response includes the output of interest and the voltage across the nonlinear device.
5. Repeat Steps 2, 3, and 4 for higher-order responses. The overall response is equal to the output components found in Steps 1, 4, etc.

The following example illustrates the procedure.

Example 2.35

Determine $H_3(\omega_1, \omega_2, \omega_3)$ for the circuit of Fig. 2.72.

Solution:

In this case, the output voltage also appears across the nonlinear device. We know that $H_1(\omega_1) = (R_1 C_0 j\omega_1 + 1)^{-1}$. Thus, with $V_{in}(t) = V_0 \exp(j\omega_1 t)$, the voltage across the capacitor is equal to

$$V_{C1}(t) = \frac{V_0}{R_1 C_0 j\omega_1 + 1} e^{j\omega_1 t}. \tag{2.239}$$

In the second step, we apply $V_{in}(t) = V_0 \exp(j\omega_1 t) + V_0 \exp(j\omega_2 t)$, obtaining the linear voltage across C_1 as

$$V_{C1}(t) = \frac{V_0 e^{j\omega_1 t}}{R_1 C_0 j\omega_1 + 1} + \frac{V_0 e^{j\omega_2 t}}{R_1 C_0 j\omega_2 + 1}. \tag{2.240}$$

With this voltage, we compute the nonlinear current flowing through C_1:

$$I_{C1,non}(t) = \alpha C_0 V_{C1} \frac{dV_{C1}}{dt} \tag{2.241}$$

$$= \alpha C_0 \left(\frac{V_0 e^{j\omega_1 t}}{R_1 C_0 j\omega_1 + 1} + \frac{V_0 e^{j\omega_2 t}}{R_1 C_0 j\omega_2 + 1} \right) \times \left(\frac{j\omega_1 V_0 e^{j\omega_1 t}}{R_1 C_0 j\omega_1 + 1} + \frac{j\omega_2 V_0 e^{j\omega_2 t}}{R_1 C_0 j\omega_2 + 1} \right). \tag{2.242}$$

Since only the component at $\omega_1 + \omega_2$ is of interest at this point, we rewrite the above expression as

$$I_{C1,non}(t) = \alpha C_0 \left[\frac{j(\omega_1 + \omega_2) V_0^2 e^{j(\omega_1+\omega_2)t}}{(R_1 C_0 j\omega_1 + 1)(R_1 C_0 j\omega_2 + 1)} + \cdots \right] \tag{2.243}$$

$$= \alpha C_0 [j(\omega_1 + \omega_2) V_0^2 e^{j(\omega_1+\omega_2)t} H_1(\omega_1) H_1(\omega_2) + \cdots]. \tag{2.244}$$

In the third step, we set the input to zero, assume a linear capacitor, and apply $I_{C1,non}(t)$ in parallel with C_1 (Fig. 2.73). The current component at $\omega_1 + \omega_2$ flows through the parallel combination of R_1 and C_0, producing $V_{C1,non}(t)$:

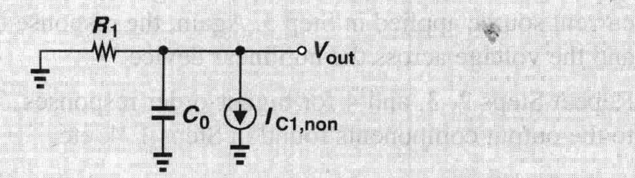


Figure 2.73 *Inclusion of nonlinear current in RC section.*

Example 2.35 (*Continued*)

$$V_{C1,non}(t) = -\alpha C_0 j(\omega_1+\omega_2)V_0^2 e^{j(\omega_1+\omega_2)t}H_1(\omega_1) \times H_1(\omega_2)\frac{R_1}{R_1C_0 j(\omega_1+\omega_2)+1} \quad (2.245)$$

$$= -\alpha R_1 C_0 j(\omega_1+\omega_2)H_1(\omega_1)H_1(\omega_2)H_1(\omega_1+\omega_2)V_0^2 e^{j(\omega_1+\omega_2)t}. \quad (2.246)$$

We note that the coefficient of $V_0^2\exp[j(\omega_1+\omega_2)t]$ in these two equations is the same as $H_2(\omega_1,\omega_2)$ in (2.229).

To determine $H_3(\omega_1,\omega_2,\omega_3)$, we must assume an input of the form $V_{in}(t) = V_0\exp(j\omega_1 t) + V_0\exp(j\omega_2 t) + V_0\exp(j\omega_3 t)$ and write the voltage across C_1 as

$$V_{C1}(t) = H_1(\omega_1)V_0e^{j\omega_1 t} + H_1(\omega_2)V_0e^{j\omega_2 t} + H_1(\omega_3)V_0e^{j\omega_3 t} + H_2(\omega_1,\omega_2)V_0^2e^{j(\omega_1+\omega_2)t} + H_2(\omega_1,\omega_3)V_0^2e^{j(\omega_1+\omega_3)t} + H_2(\omega_2,\omega_3)V_0^2e^{j(\omega_2+\omega_3)t}. \quad (2.247)$$

Note that, in contrast to Eq. (2.240), we have included the second-order nonlinear terms in the voltage so as to calculate the third-order terms.[27] The nonlinear current through C_1 is thus equal to

$$I_{C1,non}(t) = \alpha C_0 V_{C1}\frac{dV_{C1}}{dt}. \quad (2.248)$$

We substitute for V_{C1} and group the terms containing $\omega_1+\omega_2+\omega_3$:

$$I_{C1,non}(t) = \alpha C_0[H_1(\omega_1)H_2(\omega_2,\omega_3)j(\omega_2+\omega_3) + H_2(\omega_2,\omega_3)j\omega_1 H_1(\omega_1) + H_1(\omega_2)H_2(\omega_1,\omega_3)j(\omega_1+\omega_3) + H_2(\omega_1,\omega_3)j\omega_2 H_1(\omega_2) + H_1(\omega_3)H_2(\omega_1,\omega_2)j(\omega_1+\omega_2) + H_2(\omega_1,\omega_2)j\omega_3 H_1(\omega_3)]V_0^3 e^{j(\omega_1+\omega_2+\omega_3)t} + \cdots. \quad (2.249)$$

This current flows through the parallel combination of R_1 and C_0, yielding $V_{C1,non}(t)$. The reader can readily show that the coefficient of $\exp[j(\omega_1+\omega_2+\omega_3)t]$ in $V_{C1,non}(t)$ is the same as the third kernel expressed by Eq. (2.238).

The procedure described above applies to two-terminal nonlinear devices. For transistors, a similar approach can be taken. We illustrate this point with the aid of an example.

Example 2.36

Figure 2.74(a) shows the input network of a commonly-used LNA (Chapter 5). Assuming that $g_m L_1/C_{GS} = R_S$ (Chapter 5) and $I_D = \alpha(V_{GS}-V_{TH})^2$, determine the nonlinear terms in I_{out}. Neglect other capacitances, channel-length modulation, and body effect.

(*Continues*)

27. Other terms are excluded because they do not lead to a component at $\omega_1+\omega_2+\omega_3$.

Example 2.36 (*Continued*)

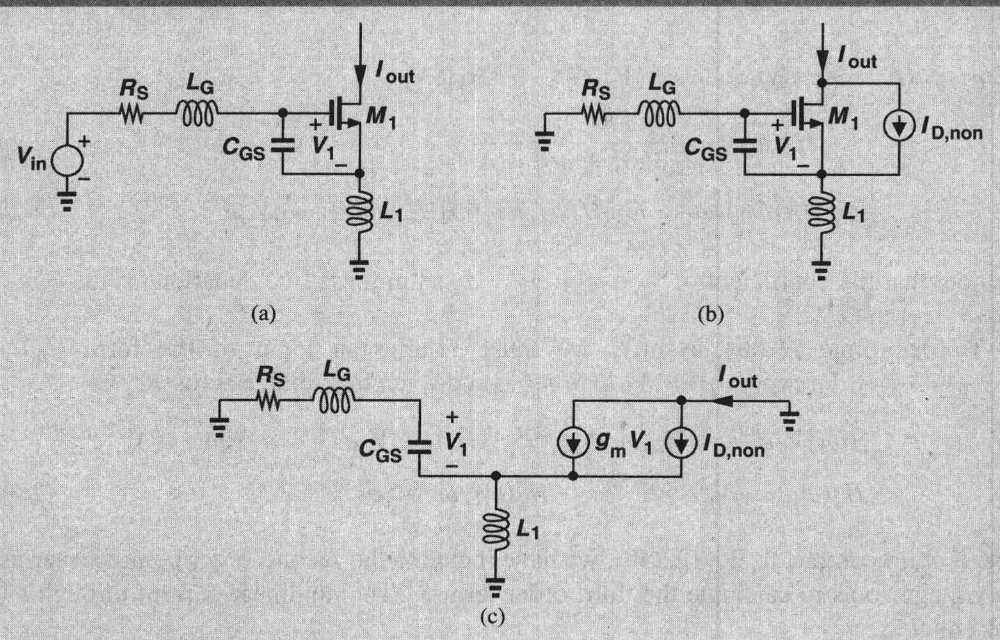


Figure 2.74 *(a) CS stage with inductors in series with source and gate, (b) inclusion of nonlinear current, (c) computation of output current.*

Solution:

In this circuit, two quantities are of interest, namely, the output current, I_{out} ($= I_D$), and the gate-source voltage, V_1; the latter must be computed each time as it determines the nonlinear component in I_D.

Let us begin with the linear response. Since the current flowing through L_1 is equal to $V_1 C_{GS} s + g_m V_1$ and that flowing through R_S and L_G equal to $V_1 C_{GS} s$, we can write a KVL around the input loop as

$$V_{in} = (R_S + L_G s)V_1 C_{GS} s + V_1 + (V_1 C_{GS} s + g_m V_1)L_1 s. \tag{2.250}$$

It follows that

$$\frac{V_1}{V_{in}} = \frac{1}{(L_1 + L_G)C_{GS}s^2 + (R_S C_{GS} + g_m L_1)s + 1}. \tag{2.251}$$

Since we have assumed $g_m L_1 / C_{GS} = R_S$, for $s = j\omega$ we obtain

$$\frac{V_1}{V_{in}}(j\omega) = \frac{1}{2g_m L_1 j\omega + 1 - \dfrac{\omega^2}{\omega_0^2}} = H_1(\omega), \tag{2.252}$$

where $\omega_0^2 = [(L_1 + L_G)C_{GS}]^{-1}$. Note that $I_{out} = g_m V_1 = g_m H_1(\omega) V_{in}$.

Example 2.36 *(Continued)*

Now, we assume $V_{in}(t) = V_0 \exp(j\omega_1 t) + V_0 \exp(j\omega_2 t)$ and write

$$V_1(t) = H_1(\omega_1)V_0 e^{j\omega_1 t} + H_1(\omega_2)V_0 e^{j\omega_2 t}. \tag{2.253}$$

Upon experiencing the characteristic $I_D = \alpha V_1^2$, this voltage results in a nonlinear current given by

$$I_{D,non} = 2\alpha H_1(\omega_1)H_1(\omega_2)V_0^2 e^{j(\omega_1+\omega_2)t}. \tag{2.254}$$

In the next step, we set V_{in} to zero and insert a current source having the above value in parallel with the drain current source [Fig. 2.74(b)]. We must compute V_1 in response to $I_{D,non}$, assuming the circuit is *linear*. From the equivalent circuit shown in Fig. 2.74(c), we have the following KVL:

$$(R_S + L_G s)V_1 C_{GS} s + V_1 + (g_m V_1 + I_{D,non} + V_1 C_{GS} s)L_1 s = 0. \tag{2.255}$$

Thus, for $s = j\omega$

$$\frac{V_1}{I_{D,non}}(j\omega) = \frac{-jL_1\omega}{2g_m L_1 j\omega + 1 - \dfrac{\omega^2}{\omega_0^2}}. \tag{2.256}$$

Since $I_{D,non}$ contains a frequency component at $\omega_1 + \omega_2$, the above transfer function must be calculated at $\omega_1 + \omega_2$ and multiplied by $I_{D,non}$ to yield V_1. We therefore have

$$H_2(\omega_1, \omega_2) = \frac{-jL_1(\omega_1+\omega_2)}{2g_m L_1 j(\omega_1+\omega_2) + 1 - \dfrac{(\omega_1+\omega_2)^2}{\omega_0^2}} 2\alpha H_1(\omega_1)H_1(\omega_2). \tag{2.257}$$

In our last step, we assume $V_{in}(t) = V_0 \exp(j\omega_1 t) + V_0 \exp(j\omega_2 t) + V_0 \exp(j\omega_3 t)$ and write

$$V_1(t) = H_1(\omega_1)V_0 e^{j\omega_1 t} + H_1(\omega_2)V_0 e^{j\omega_2 t} + H_1(\omega_3)V_0 e^{j\omega_3 t} + H_2(\omega_1, \omega_2)V_0^2 e^{j(\omega_1+\omega_2)t}$$
$$+ H_2(\omega_1, \omega_3)V_0^2 e^{j(\omega_1+\omega_3)t} + H_2(\omega_2, \omega_3)V_0^2 e^{j(\omega_2+\omega_3)t}. \tag{2.258}$$

Since $I_D = \alpha V_1^2$, the nonlinear current at $\omega_1 + \omega_2 + \omega_3$ is expressed as

$$I_{D,non} = 2\alpha[H_1(\omega_1)H_2(\omega_2, \omega_3) + H_1(\omega_2)H_2(\omega_1, \omega_3)$$
$$+ H_1(\omega_3)H_2(\omega_1, \omega_2)]V_0^3 e^{j(\omega_1+\omega_2+\omega_3)t}. \tag{2.259}$$

The third-order nonlinear component in the output of interest, I_{out}, is equal to the above expression. We note that, even though the transistor exhibits only second-order nonlinearity, the degeneration (feedback) caused by L_1 results in higher-order terms.

The reader is encouraged to repeat this analysis using the harmonic method and see that it is much more complex.

REFERENCES

[1] B. Razavi, *Design of Analog CMOS Integrated Circuits,* Boston: McGraw-Hill, 2001.

[2] L. W. Couch, *Digital and Analog Communication Systems,* Fourth Edition, New York: Macmillan Co., 1993.

[3] A. van der Ziel, "Thermal Noise in Field Effect Transistors," *Proc. IRE,* vol. 50, pp. 1808–1812, Aug. 1962.

[4] A. A. Abidi, "High-Frequency Noise Measurements on FETs with Small Dimensions," *IEEE Trans. Electron Devices,* vol. 33, pp. 1801–1805, Nov. 1986.

[5] A. J. Sholten et al., "Accurate Thermal Noise Model of Deep-Submicron CMOS," *IEDM Dig. Tech. Papers,* pp. 155–158, Dec. 1999.

[6] B. Razavi, "Impact of Distributed Gate Resistance on the Performance of MOS Devices," *IEEE Trans. Circuits and Systems- Part I,* vol. 41, pp. 750–754, Nov. 1994.

[7] H. T. Friis, "Noise Figure of Radio Receivers," *Proc. IRE,* vol. 32, pp. 419–422, July 1944.

[8] A. Papoulis, *Probability, Random Variables, and Stochastic Processes,* Third Edition, New York: McGraw-Hill, 1991.

[9] R. W. Bennet, "Methods of Solving Noise Problems," *Proc. IRE,* vol. 44, pp. 609–638, May 1956.

[10] S. Narayanan, "Application of Volterra Series to Intermodulation Distortion Analysis of Transistor Feedback Amplifiers," *IEEE Tran. Circuit Theory,* vol. 17, pp. 518–527, Nov. 1970.

[11] P. Wambacq et al., "High-Frequency Distortion Analysis of Analog Integrated Circuits," *IEEE Tran. Circuits and Systems, II,* vol. 46, pp. 335–334, March 1999.

[12] P. Wambaq and W. Sansen, *Distortion Analysis of Analog Integrated Circuits,* Norwell, MA: Kluwer, 1998.

[13] J. Bussganag, L. Ehrman, and J. W. Graham, "Analysis of Nonlinear Systems with Multiple Inputs," *Proc. IEEE,* vol. 62, pp. 1088–1119, Aug. 1974.

[14] E. Bedrosian and S. O. Rice, "The Output Properties of Volterra Systems (Nonlinear Systems with Memory) Driven by Harmonic and Gaussian Inputs," *Proc. IEEE,* vol. 59, pp. 1688–1707, Dec. 1971.

PROBLEMS

2.1. Two nonlinear stages are cascaded. If the input/output characteristic of each stage is approximated by a third-order polynomial, determine the P_{1dB} of the cascade in terms of the P_{1dB} of each stage.

2.2. Repeat Example 2.11 if one interferer has a level of -3 dBm and the other, -35 dBm.

2.3. If cascaded, stages having only *second-order* nonlinearity can yield a finite IP_3. For example, consider the cascade identical common-source stages shown in Fig. 2.75.

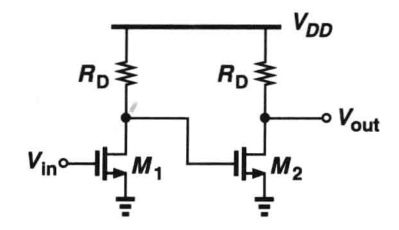


Figure 2.75 *Cascade of CS stages.*

If each transistor operates in saturation and follows the ideal square-law behavior, determine the IP_3 of the cascade.

2.4. Determine the IP_3 and P_{1dB} for a system whose characteristic is approximated by a fifth-order polynomial.

2.5. Consider the scenario shown in Fig. 2.76, where $\omega_3 - \omega_2 = \omega_2 - \omega_3$ and the band-pass filter provides an attenuation of 17 dB at ω_2 and 37 dB at ω_3.

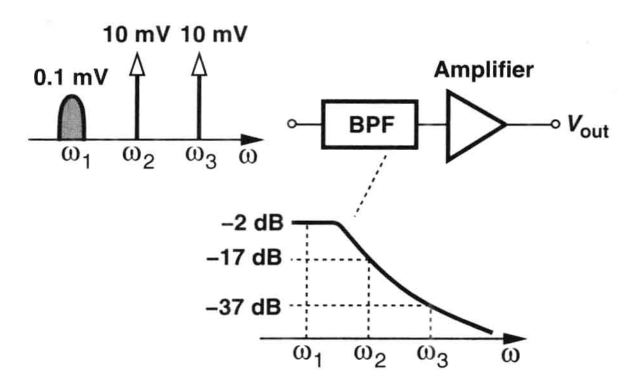


Figure 2.76 *Cascade of BPF and amplifier.*

(a) Compute the IIP_3 of the amplifier such that the intermodulation product falling at ω_1 is 20 dB below the desired signal.

(b) Suppose an amplifier with a voltage gain of 10 dB and $IIP_3 = 500\,\text{mV}_p$ precedes the band-pass filter. Calculate the IIP_3 of the overall chain. (Neglect second-order nonlinearities.)

2.6. Prove that the Fourier transform of the autocorrelation of a random signal yields the spectrum, i.e., the power measured in a 1-Hz bandwidth at each frequency.

2.7. A broadband circuit sensing an input $V_0 \cos \omega_0 t$ produces a third harmonic $V_3 \cos(3\omega_0 t)$. Determine the 1-dB compression point in terms of V_0 and V_3.

2.8. Prove that in Fig. 2.36, the noise power delivered by R_1 to R_2 is equal to that delivered by R_2 to R_1 if the resistors reside at the same temperature. What happens if they do not?

2.9. Explain why the channel thermal noise of a MOSFET is modeled by a current source tied between the source and drain terminals (rather than, say, between the gate and source terminals).

2.10. Prove that the channel thermal noise of a MOSFET can be referred to the gate as a voltage given by $4kT\gamma/g_m$. As shown in Fig. 2.77, the two circuits must generate the same current with the same terminal voltages.

2.11. Determine the NF of the circuit shown in Fig. 2.52 using Friis' equation.

2.12. Prove that the output noise voltage of the circuit shown in Fig. 2.46(c) is given by $\overline{V_{n2}^2} = \overline{I_{n1}^2} r_O^2$.

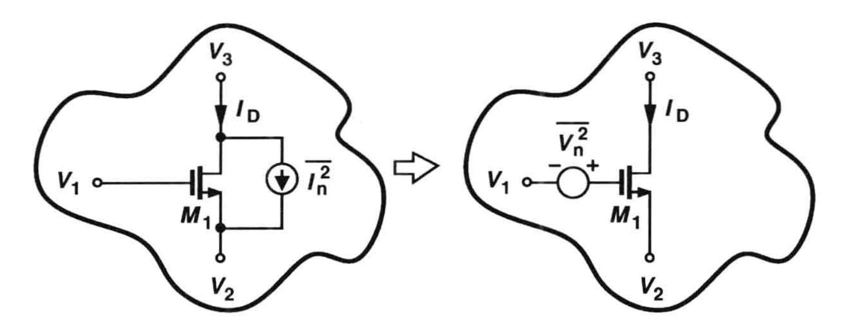


Figure 2.77 *Equivalent circuits for noise of a MOSFET.*

2.13. Repeat Example 2.23 if the CS and CG stages are swapped. Does the NF change? Why?

2.14. Repeat Example 2.23 if R_{D1} and R_{D2} are replaced with ideal current sources and channel-length modulation is not neglected.

2.15. The noise figure of a circuit is known for a source impedance of R_{S1}. Is it possible to compute the noise figure for another source impedance R_{S2}? Explain in detail.

2.16. Repeat Example 2.21 for the arrangement shown in Fig. 2.78, where the transformer amplifies its primary voltage by a factor of n and transforms R_S to a value of n^2R_S.

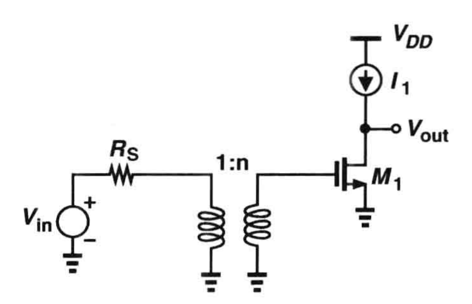


Figure 2.78 *CS stage driven by a transformer.*

2.17. For matched inputs and outputs, prove that the NF of a passive (reciprocal) circuit is equal to its power loss.

2.18. Determine the noise figure of each circuit in Fig. 2.79 with respect to a source impedance R_S. Neglect channel-length modulation and body effect.

2.19. Determine the noise figure of each circuit in Fig. 2.80 with respect to a source impedance R_S. Neglect channel-length modulation and body effect.

2.20. Determine the noise figure of each circuit in Fig. 2.81 with respect to a source impedance R_S. Neglect channel-length modulation and body effect.

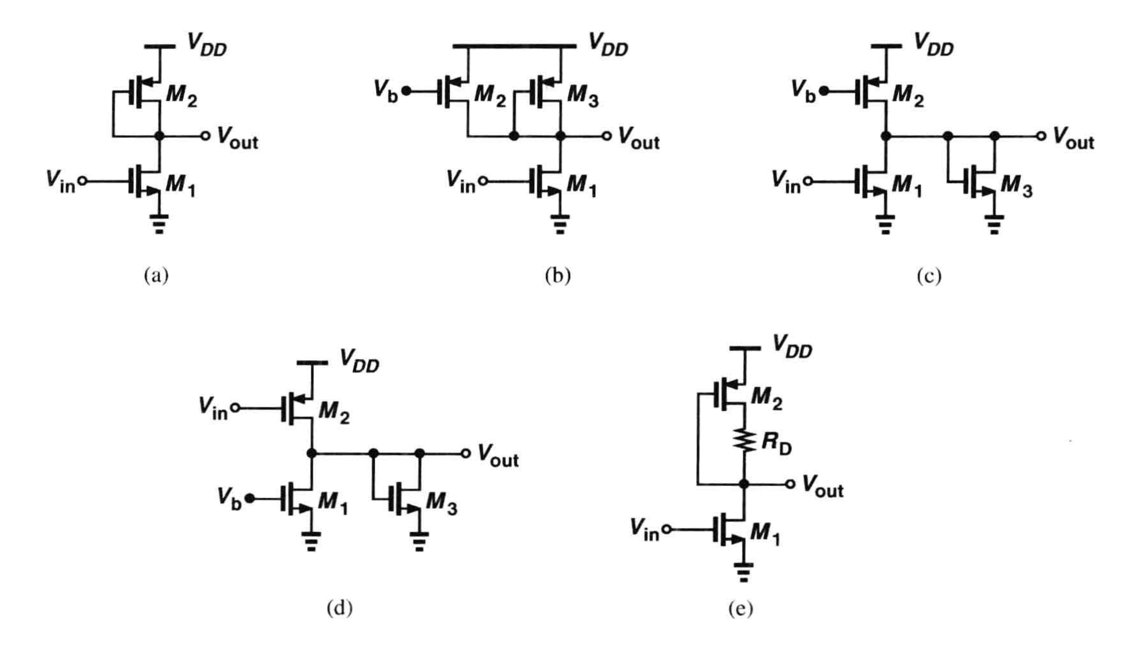


Figure 2.79 *CS stages for NF calculation.*

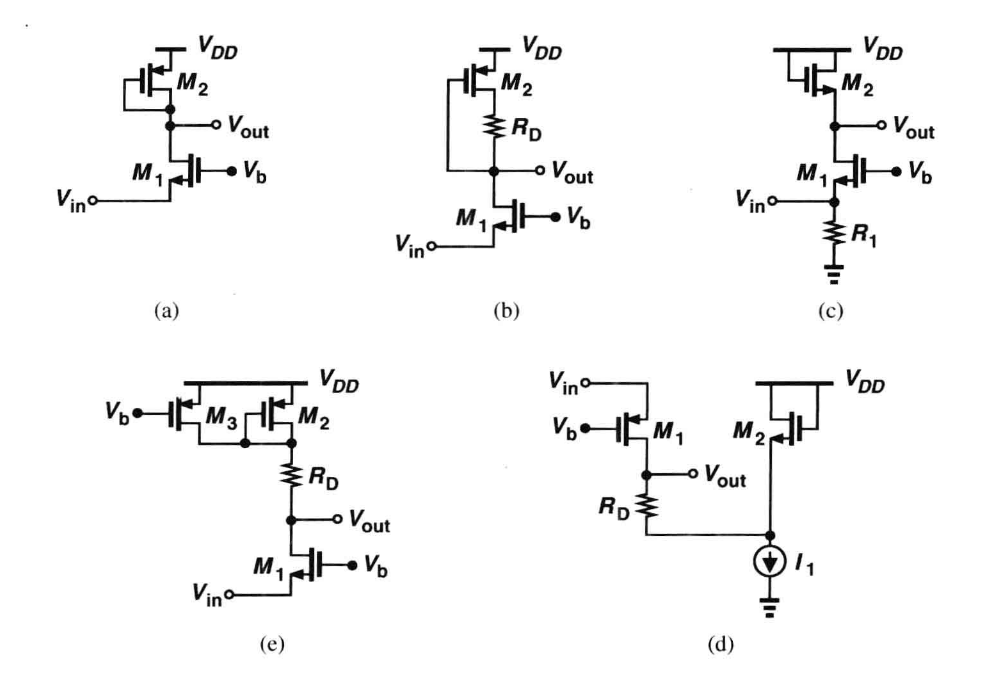


Figure 2.80 *CG stages for NF calculation.*

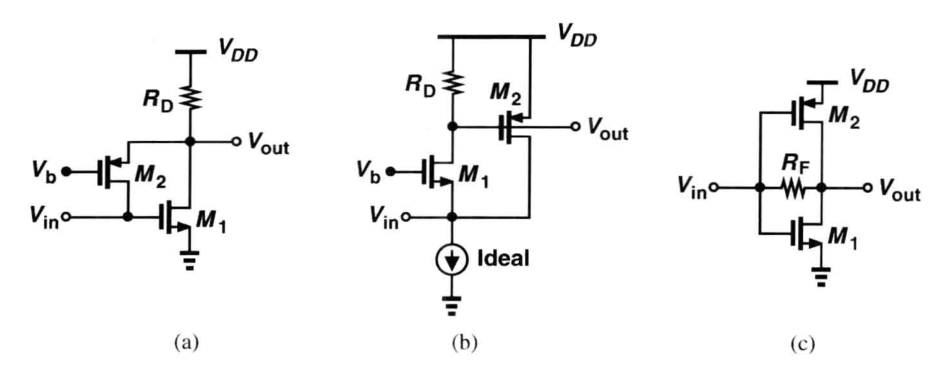


Figure 2.81 *Stages for NF calculation.*

CHAPTER



COMMUNICATION CONCEPTS

The design of highly-integrated RF transceivers requires a solid understanding of communication theory. For example, as mentioned in Chapter 2, the receiver sensitivity depends on the minimum acceptable signal-to-noise ratio, which itself depends on the type of modulation. In fact, today we rarely design a low-noise amplifier, an oscillator, etc., with no attention to the type of transceiver in which they are used. Furthermore, modern RF designers must regularly interact with digital signal processing engineers to trade functions and specifications and must therefore speak the same language.

This chapter provides a basic, yet necessary, understanding of modulation theory and wireless standards. Tailored to a reader who is ultimately interested in RF IC design rather than communication theory, the concepts are described in an intuitive language so that they can be incorporated in the reader's daily work. The outline of the chapter is shown below.

Modulation	Mobile Systems	Multiple Access Technqiues	Wireless Standards
▪ AM, PM, FM	▪ Cellular System	▪ Duplexing	▪ GSM
▪ Intersymbol Interference	▪ Hand-off	▪ FDMA	▪ IS–95 CDMA
▪ Signal Constellations	▪ Multipath Fading	▪ TDMA	▪ Wideband CDMA
▪ ASK, PSK, FSK	▪ Diversity	▪ CDMA	▪ Bluetooth
▪ QPSK, GMSK, QAM			▪ IEEE802.11a/b/g
▪ OFDM			
▪ Spectral Regrowth			

3.1 GENERAL CONSIDERATIONS

How does your voice enter a cell phone here and come out of another cell phone miles away? We wish to understand the incredible journey that your voice signal takes.

The transmitter in a cell phone must convert the voice, which is called a "baseband signal" because its spectrum (20 Hz to 20 kHz) is centered around zero frequency, to a "passband signal," i.e., one residing around a nonzero center frequency, ω_c [Fig. 3.1(b)]. We call ω_c the "carrier frequency."

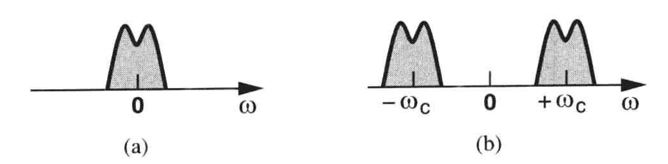


Figure 3.1 *(a) Baseband and (b) passband signal spectra.*

More generally, "modulation" converts a baseband signal to a passband signal. From another point of view, modulation varies certain parameters of a sinusoidal carrier according to the baseband signal. For example, if the carrier is expressed as $A_0 \cos \omega_c t$, then the modulated signal is given by

$$x(t) = a(t) \cos[\omega_c t + \theta(t)], \tag{3.1}$$

where the amplitude, $a(t)$ and the phase, $\theta(t)$, are modulated.

The inverse of modulation is demodulation or detection, with the goal being to reconstruct the original baseband signal with minimal noise, distortion, etc. Thus, as depicted in Fig. 3.2, a simple communication system consists of a modulator/transmitter, a channel (e.g., air or a cable), and a receiver/demodulator. Note that the channel attenuates the signal. A "transceiver" contains both a modulator and a demodulator; the two are called a "modem."

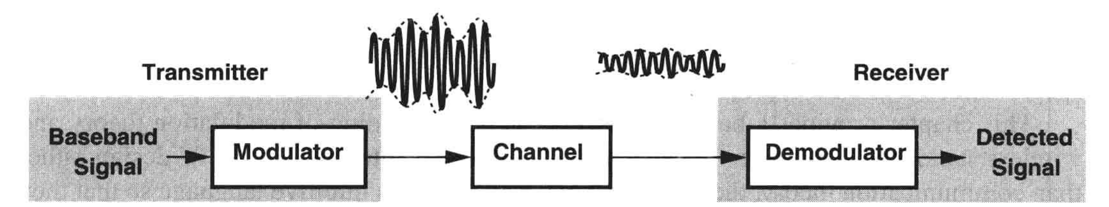


Figure 3.2 *Generic communication system.*

Important Aspects of Modulation Among various attributes of each modulation scheme, three prove particularly critical in RF design.

1. Detectability, i.e., the quality of the demodulated signal for a given amount of channel attenuation and receiver noise. As an example, consider the binary amplitude modulation shown in Fig. 3.3(a), where logical ONEs are represented by full amplitude and ZEROs by zero amplitude. The demodulation must simply distinguish between these two amplitude values. Now, suppose we wish to carry more information and hence employ four different amplitudes as depicted in Fig. 3.3(b).

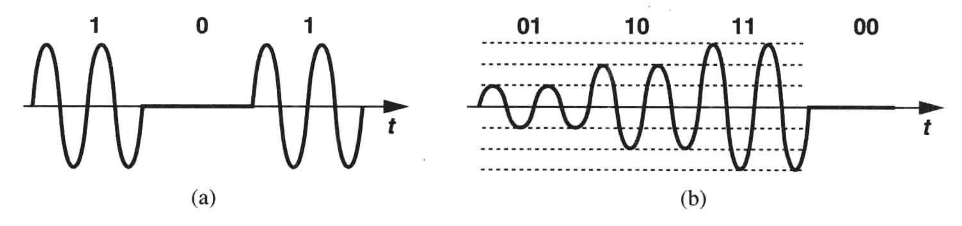


Figure 3.3 *(a) Two-level and (b) four-level modulation schemes.*

In this case, the four amplitude values are closer to one another and can therefore be misinterpreted in the presence of noise. We say the latter signal is less detectable.

2. Bandwidth efficiency, i.e., the bandwidth occupied by the modulated carrier for a given information rate in the baseband signal. This aspect plays a critical role in today's systems because the available spectrum is limited. For example, the GSM phone system provides a total bandwidth of 25 MHz for millions of users in crowded cities. The sharing of this bandwidth among so many users is explained in Section 3.6.
3. Power efficiency, i.e., the type of power amplifier (PA) that can be used in the transmitter. As explained later in this chapter, some modulated waveforms can be processed by means of *nonlinear* power amplifiers, whereas some others require linear amplifiers. Since nonlinear PAs are generally more efficient (Chapter 12), it is desirable to employ a modulation scheme that lends itself to nonlinear amplification.

The above three attributes typically trade with one another. For example, we may suspect that the modulation format in Fig. 3.3(b) is more bandwidth-efficient than that in Fig. 3.3(a) because it carries twice as much information for the same bandwidth. This advantage comes at the cost of detectability—because the amplitude values are more closely spaced—and power efficiency—because PA nonlinearity compresses the larger amplitudes.

3.2 ANALOG MODULATION

If an analog signal, e.g., that produced by a microphone, is impressed on a carrier, then we say we have performed analog modulation. While uncommon in today's high-performance communications, analog modulation provides fundamental concepts that prove essential in studying digital modulation as well.

3.2.1 Amplitude Modulation

For a baseband signal $x_{BB}(t)$, an amplitude-modulated (AM) waveform can be constructed as

$$x_{AM}(t) = A_c[1 + m x_{BB}(t)] \cos \omega_c t, \tag{3.2}$$

where m is called the "modulation index."[1] Illustrated in Fig. 3.4(a) is a multiplication method for generating an AM signal. We say the baseband signal is "upconverted." The waveform $A_c \cos \omega_c t$ is generated by a "local oscillator" (LO). Multiplication by $\cos \omega_c t$ in the time domain simply translates the spectrum of $x_{BB}(t)$ to a center frequency of ω_c [Fig. 3.4(b)]. Thus, the bandwidth of $x_{AM}(t)$ is twice that of $x_{BB}(t)$. Note that since $x_{BB}(t)$ has a symmetric spectrum around zero (because it is a real signal), the spectrum of $x_{AM}(t)$ is also symmetric around ω_c. This symmetry does not hold for all modulation schemes and plays a significant role in the design of transceiver architectures (Chapter 4).

1. Note that m has a dimension of 1/volt if $x_{BB}(t)$ is a voltage quantity.

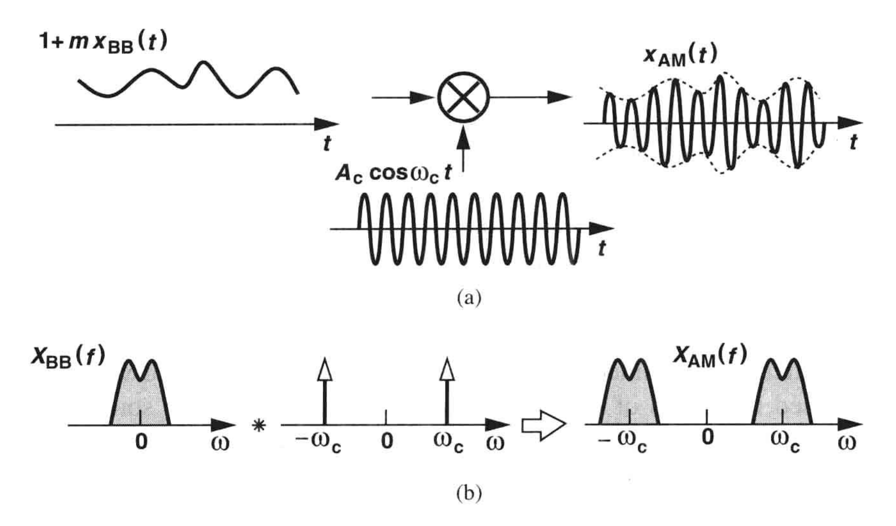


Figure 3.4 *(a) Generation of AM signal, (b) resulting spectra.*

Example 3.1

The modulated signal of Fig. 3.3(a) can be considered as the product of a random binary sequence toggling between zero and 1 and a sinusoidal carrier. Determine the spectrum of the signal.

Solution:

The spectrum of a random binary sequence with equal probabilities of ONEs and ZEROs is given by (Section 3.3.1):

$$S(f) = T_b \left(\frac{\sin \pi f T_b}{\pi f T_b} \right)^2 + 0.5\delta(f). \tag{3.3}$$

Multiplication by a sinusoid in the time domain shifts this spectrum to a center frequency of $\pm f_c$ (Fig. 3.5).

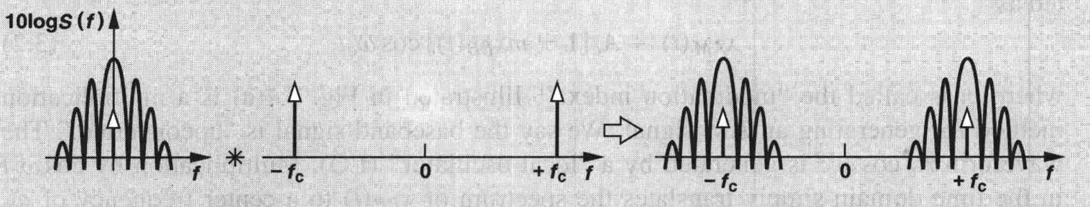


Figure 3.5 *Spectrum of random binary data and AM output.*

Except for broadcast AM radios, amplitude modulation finds limited use in today's wireless systems. This is because carrying analog information in the amplitude requires a highly-linear power amplifier in the transmitter. Amplitude modulation is also more sensitive to additive noise than phase or frequency modulation is.

3.2.2 Phase and Frequency Modulation

Phase modulation (PM) and frequency modulation (FM) are important concepts that are encountered not only within the context of modems but also in the analysis of such circuits as oscillators and frequency synthesizers.

Let us consider Eq. (3.1) again. We call the argument $\omega_c t + \theta(t)$ the "total phase." We also define the "instantaneous frequency" as the time derivative of the phase; thus, $\omega_c + d\theta/dt$ is the "total frequency" and $d\theta/dt$ is the "excess frequency" or the "frequency deviation." If the amplitude is constant and the excess phase is linearly proportional to the baseband signal, we say the carrier is phase-modulated:

$$x_{PM}(t) = A_c \cos[\omega_c t + m x_{BB}(t)], \tag{3.4}$$

where m denotes the phase modulation index. To understand PM intuitively, first note that, if $x_{BB}(t) = 0$, then the zero-crossing points of the carrier waveform occur at uniformly-spaced instants equal to integer multiples of the period, $T_c = 1/\omega_c$. For a time-varying $x_{BB}(t)$, on the other hand, the zero crossings are modulated (Fig. 3.6) while the amplitude remains constant.

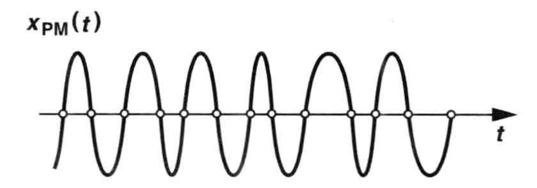


Figure 3.6 *Zero crossings in a phase-modulated signal.*

Similarly, if the excess frequency, $d\theta/dt$, is linearly proportional to the baseband signal, we say the carrier is frequency-modulated:

$$x_{FM}(t) = A_c \cos[\omega_c t + m \int_{-\infty}^{t} x_{BB}(\tau) d\tau]. \tag{3.5}$$

Note that the instantaneous frequency is equal to $\omega_c + m x_{BB}(t)$.[2]

Example 3.2

Determine the PM and FM signals in response to (a) $x_{BB}(t) = A_0$, (b) $x_{BB}(t) = \alpha t$.

Solution:

(a) For a constant baseband signal,

$$x_{PM}(t) = A_c \cos(\omega_c t + m A_0); \tag{3.6}$$

(Continues)

2. In this case, m has a dimension of radian frequency/volt if $x_{BB}(t)$ is a voltage quantity.

Example 3.2 *(Continued)*

i.e., the PM output simply contains a constant phase shift. The FM output, on the other hand, is expressed as

$$x_{FM}(t) = A_c \cos(\omega_c t + mA_0 t) \tag{3.7}$$

$$= A_c \cos[(\omega_c + mA_0)t]. \tag{3.8}$$

Thus, the FM output exhibits a constant frequency shift equal to mA_0.
(b) If $x_{BB}(t) = \alpha t$, then

$$x_{PM}(t) = A_c \cos(\omega_c t + m\alpha t) \tag{3.9}$$

$$= A_c \cos[(\omega_c + m\alpha)t], \tag{3.10}$$

i.e., the PM output experiences a constant frequency shift. For the FM output, we have,

$$x_{FM}(t) = A_c \cos\left(\omega_c t + \frac{m\alpha}{2} t^2\right). \tag{3.11}$$

This signal can be viewed as a waveform whose phase grows quadratically with time.

The nonlinear dependence of $x_{PM}(t)$ and $x_{FM}(t)$ upon $x_{BB}(t)$ generally *increases* the occupied bandwidth. For example, if $x_{BB}(t) = A_m \cos \omega_m t$, then

$$x_{FM}(t) = A_c \cos\left(\omega_c t + \frac{mA_m}{\omega_m} \sin \omega_m t\right), \tag{3.12}$$

exhibiting spectral lines well beyond $\omega_c \pm \omega_m$. Various approximations for the bandwidth of PM and FM signals have been derived [1–3].

Narrowband FM Approximation A special case of FM that proves useful in the analysis of RF circuits and systems arises if $mA_m/\omega_m \ll 1$ rad in Eq. (3.12). The signal can then be approximated as

$$x_{FM}(t) \approx A_c \cos \omega_c t - A_m A_c \frac{m}{\omega_m} \sin \omega_m t \sin \omega_c t \tag{3.13}$$

$$\approx A_c \cos \omega_c t - \frac{mA_m A_c}{2\omega_m} \cos(\omega_c - \omega_m)t + \frac{mA_m A_c}{2\omega_m} \cos(\omega_c + \omega_m)t. \tag{3.14}$$

Illustrated in Fig. 3.7, the spectrum consists of impulses at $\pm\omega_c$ (the carrier) and "sidebands" at $\omega_c \pm \omega_m$ and $-\omega_c \pm \omega_m$. Note that, as the modulating frequency, ω_m, increases, the magnitude of the sidebands decreases.

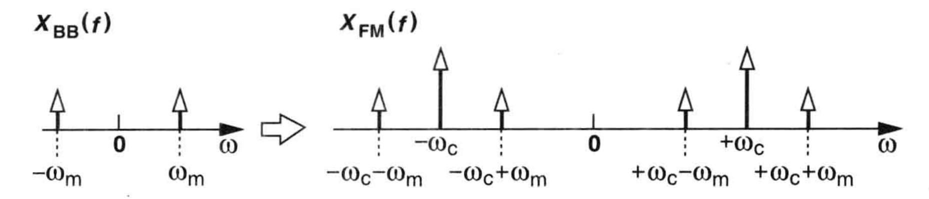


Figure 3.7 *Spectrum of a narrowband FM signal.*

Example 3.3

It is sometimes said that the FM (or PM) sidebands have opposite signs, whereas AM sidebands have identical signs. Is this generally true?

Solution:

Equation (3.14) indeed suggests that $\cos(\omega_c - \omega_m)t$ and $\cos(\omega_c + \omega_m)t$ have opposite signs. Figure 3.8(a) illustrates this case by allowing signs in the magnitude plot. For a carrier whose amplitude is modulated by a sinusoid, we have

$$x_{AM}(t) = A_c(1 + m \cos \omega_m t) \cos \omega_c t \tag{3.15}$$

$$= A_c \cos \omega_c t + \frac{mA_c}{2} \cos(\omega_c + \omega_m)t + \frac{mA_c}{2} \cos(\omega_c - \omega_m)t. \tag{3.16}$$

Thus, it appears that the sidebands have identical signs [Fig. 3.8(b)]. However, in general, the polarity of the sidebands per se does not distinguish AM from FM. Writing the four possible combinations of sine and cosine in Eqs. (3.2) and (3.5), the reader can arrive at the spectra shown in Fig. 3.9. Given the exact waveforms for the carrier and the sidebands, one can decide from these spectra whether the modulation is AM or narrowband FM.

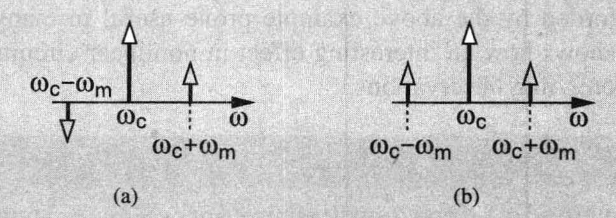


Figure 3.8 *Spectra of (a) narrowband FM and (b) AM signals.*

However, an important difference between the AM and FM sidebands relates to their angular rotation with respect to the carrier. In an AM signal, the sidebands must modulate only the amplitude at any time. Thus, as illustrated by phasors in Fig. 3.10(a), the two must rotate in opposite directions such that their resultant remains *aligned* with the carrier. On the other hand, the sidebands of an FM signal must create no component along the carrier amplitude, and hence are positioned as shown in Fig. 3.10(b) so that their resultant remains *perpendicular* to the carrier at all times.

(*Continues*)

Example 3.3 (*Continued*)

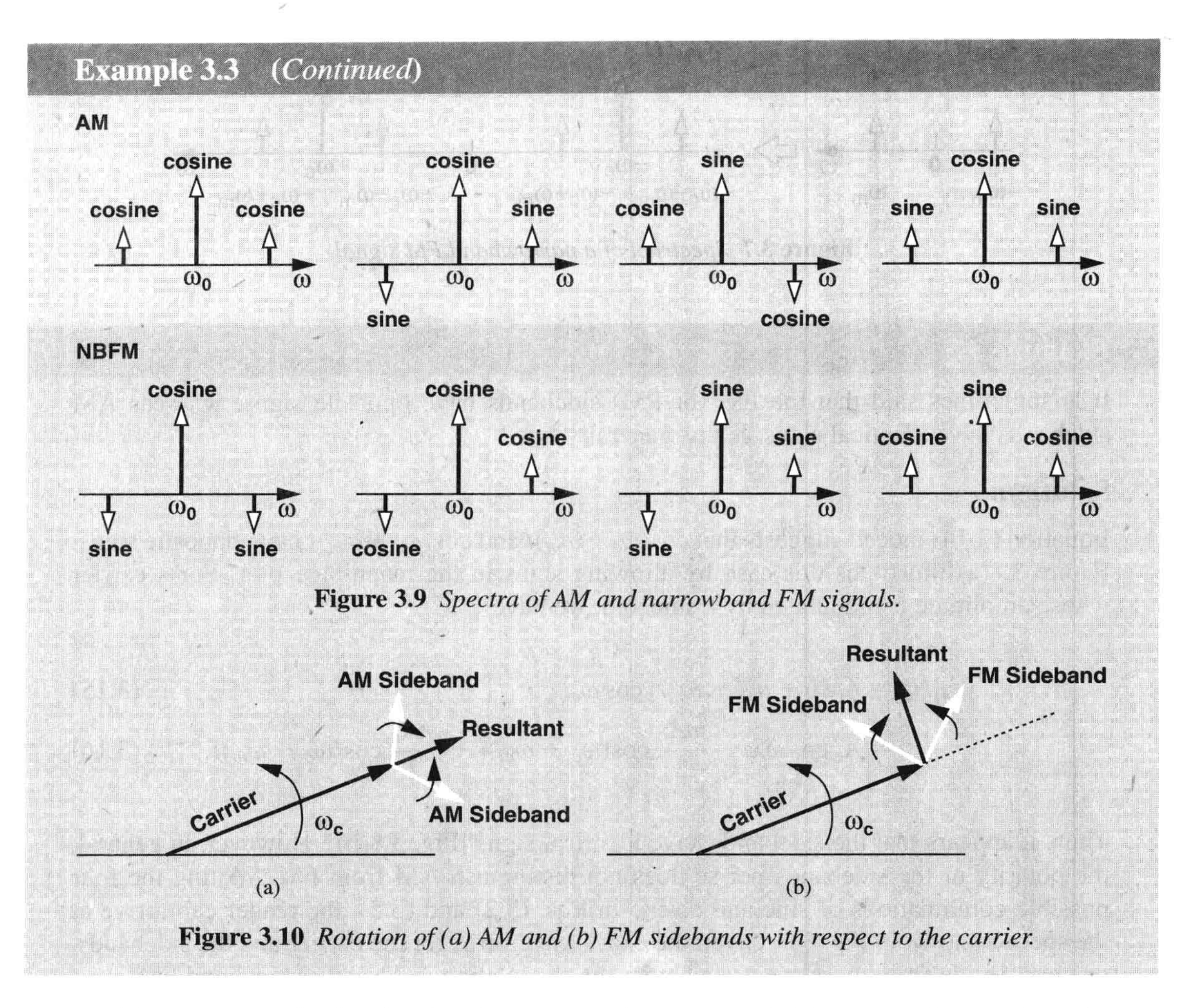


Figure 3.9 *Spectra of AM and narrowband FM signals.*

Figure 3.10 *Rotation of (a) AM and (b) FM sidebands with respect to the carrier.*

The insights afforded by the above example prove useful in many RF circuits. The following example shows how an interesting effect in nonlinear circuits can be explained with the aid of the foregoing observations.

Example 3.4

The sum of a large sinusoid at ω_c and a small sinusoid at $\omega_c + \omega_m$ is applied to a differential pair [Fig. 3.11(a)]. Explain why the output spectrum contains a component at $\omega_c - \omega_m$. Assume that the differential pair experiences "hard limiting," i.e., A is large enough to steer I_{SS} to each side.

Solution:

Let us decompose the input spectrum into two symmetric spectra as shown in Fig. 3.11(b).[3] The one with sidebands of identical signs can be viewed as an AM waveform, which, due to hard limiting, is suppressed at the output. The spectrum with sidebands of opposite

3. We call these symmetric because omission of sideband signs would make them symmetric.

Example 3.4 (*Continued*)

signs can be considered an FM waveform, which emerges at the output intact because hard limiting does not affect the zero crossings of the waveform.

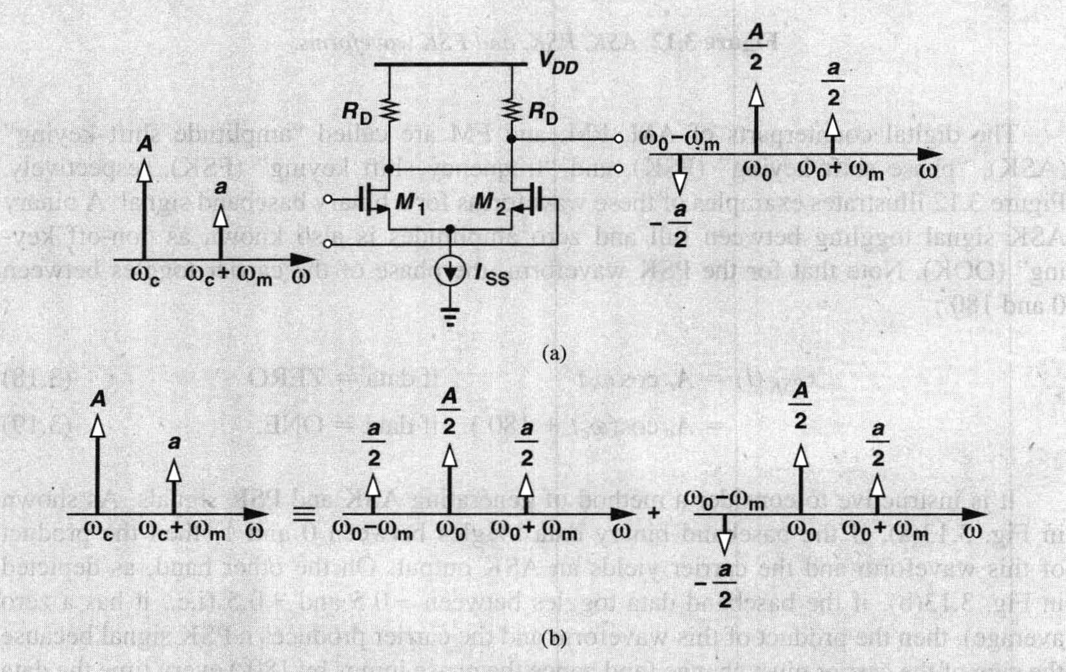


Figure 3.11 *(a) Differential pair sensing a large and a small signal, (b) conversion of one sideband to AM and FM components.*

The reader may wonder how we decided that the two symmetric spectra in Fig. 3.11(b) are AM and FM, respectively. We write the inputs in the time domain as

$$\begin{aligned} A\cos\omega_c t + a\cos(\omega_c+\omega_m)t &= \frac{A}{2}\cos\omega_c t + \frac{a}{2}\cos(\omega_c+\omega_m)t + \frac{a}{2}\cos(\omega_c-\omega_m)t \\ &\quad + \frac{A}{2}\cos\omega_c t + \frac{a}{2}\cos(\omega_c+\omega_m)t \\ &\quad - \frac{a}{2}\cos(\omega_c-\omega_m)t. \end{aligned} \tag{3.17}$$

Based on the observations in Example 3.3, we recognize the first three terms in Eq. (3.17) as an AM signal and the last three terms as an FM signal.

3.3 DIGITAL MODULATION

In digital communication systems, the carrier is modulated by a digital baseband signal. For example, the voice produced by the microphone in a cell phone is digitized and subsequently impressed on the carrier. As explained later in this chapter, carrying the information in digital form offers many advantages over communication in the analog domain.

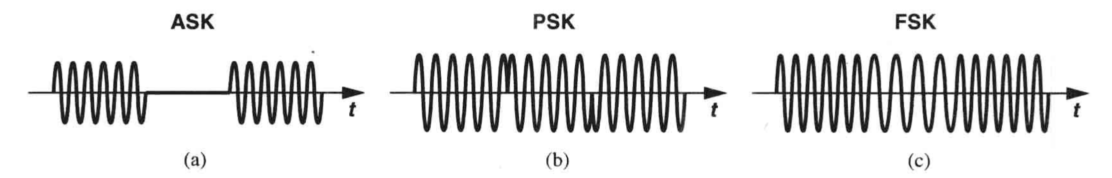


Figure 3.12 *ASK, PSK, and FSK waveforms.*

The digital counterparts of AM, PM, and FM are called "amplitude shift keying" (ASK), "phase shift keying" (PSK), and "frequency shift keying" (FSK), respectively. Figure 3.12 illustrates examples of these waveforms for a binary baseband signal. A binary ASK signal toggling between full and zero amplitudes is also known as "on-off keying" (OOK). Note that for the PSK waveform, the phase of the carrier toggles between 0 and 180°:

$$x_{PSK}(t) = A_c \cos \omega_c t \qquad \text{if data} = \text{ZERO} \tag{3.18}$$

$$= A_c \cos(\omega_c t + 180^\circ) \quad \text{if data} = \text{ONE.} \tag{3.19}$$

It is instructive to consider a method of generating ASK and PSK signals. As shown in Fig. 3.13(a), if the baseband binary data toggles between 0 and 1, then the product of this waveform and the carrier yields an ASK output. On the other hand, as depicted in Fig. 3.13(b), if the baseband data toggles between −0.5 and +0.5 (i.e., it has a zero average), then the product of this waveform and the carrier produces a PSK signal because the sign of the carrier must change (and hence the phase jumps by 180°) every time the data changes.

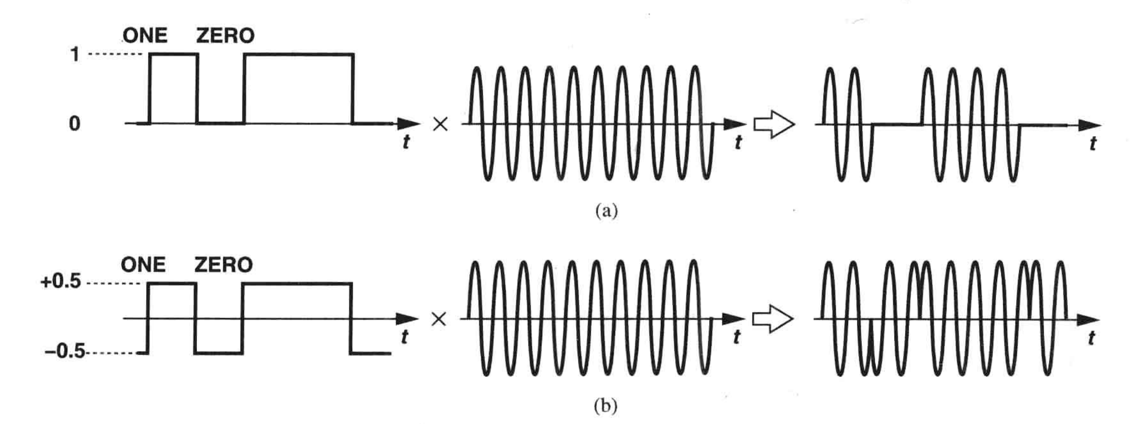


Figure 3.13 *Generation of (a) ASK and (b) PSK signals.*

In addition to ASK, PSK, and FSK, numerous other digital modulation schemes have been introduced. In this section, we study those that find wide application in RF systems. But, we must first familiarize ourselves with two basic concepts in digital communications: "intersymbol interference" (ISI) and "signal constellations."

3.3.1 Intersymbol Interference

Linear time-invariant systems can "distort" a signal if they do not provide sufficient bandwidth. A familiar example of such a behavior is the attenuation of high-frequency components of a periodic square wave in a low-pass filter [Fig. 3.14(a)]. However, limited bandwidth more detrimentally impacts *random* bit streams. To understand the issue, first recall that if a single ideal rectangular pulse is applied to a low-pass filter, then the output exhibits an exponential tail that becomes longer as the filter bandwidth decreases. This occurs fundamentally because a signal cannot be both time-limited and bandwidth-limited: when the time-limited pulse passes through the band-limited system, the output must extend to infinity in the time domain.

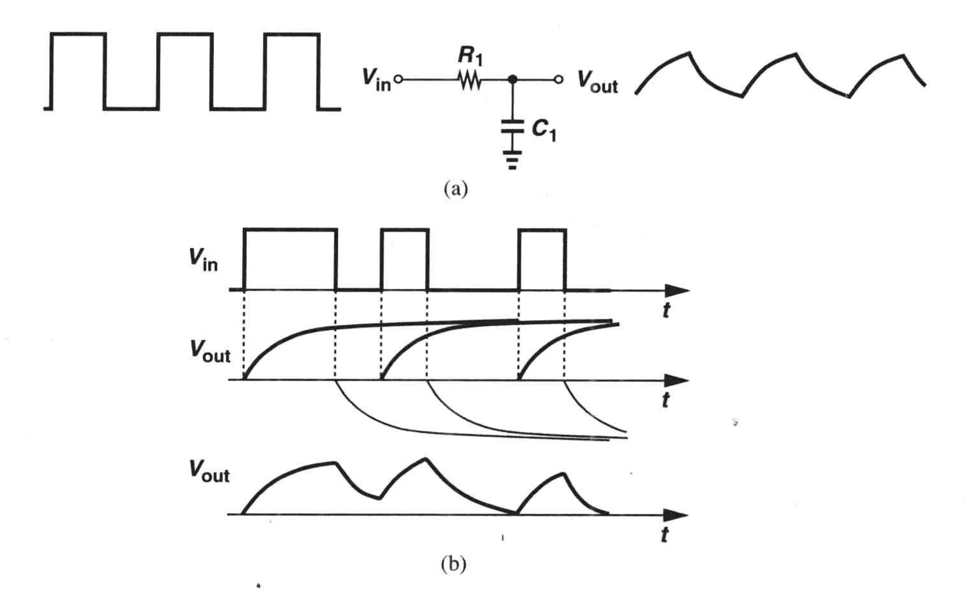


Figure 3.14 *Effect of low-pass filter on (a) periodic waveform and (b) random sequence.*

Now suppose the output of a digital system consists of a random sequence of ONEs and ZEROs. If this sequence is applied to a low-pass filter (LPF), the output can be obtained as the superposition of the responses to each input bit [Fig. 3.14(b)]. We note that each bit level is corrupted by decaying tails created by previous bits. Called "intersymbol interference" (ISI), this phenomenon leads to a higher error rate because it brings the peak levels of ONEs and ZEROs closer to the detection threshold. We also observe a trade-off between noise and ISI: if the bandwidth is reduced so as to lessen the integrated noise, then ISI increases.

In general, any system that removes part of the spectrum of a signal introduces ISI. This can be better seen by an example.

Example 3.5

Determine the spectrum of the random binary sequence, $x_{BB}(t)$, in Fig. 3.15(a) and explain, in the frequency domain, the effect of low-pass filtering it.

(Continues)

Example 3.5 *(Continued)*

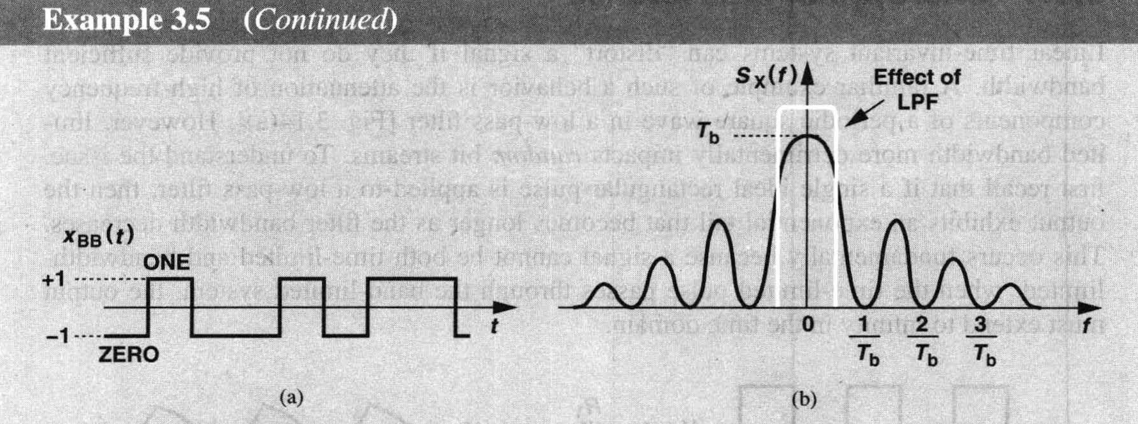


Figure 3.15 *(a) A random binary sequence toggling between −1 and +1, (b) its spectrum.*

Solution:

Consider a general random binary sequence in which the basic pulse is denoted by $p(t)$. We can express the sequence as

$$x_{BB}(t) = \sum_{n=0}^{\infty} a_n p(t - nT_b), \tag{3.20}$$

where a_n assumes a random value of +1 or −1 with equal probabilities. In this example, $p(t)$ is simply a rectangular pulse. It can be proved [1] that the spectrum of $x_{BB}(t)$ is given by the square of the magnitude of the Fourier transform of $p(t)$:

$$S_x(f) = \frac{1}{T_b}|P(f)|^2. \tag{3.21}$$

For a rectangular pulse of width T_b (and unity height),

$$P(f) = T_b \frac{\sin \pi f T_b}{\pi f T_b}, \tag{3.22}$$

yielding

$$S_x(f) = T_b \left(\frac{\sin \pi f T_b}{\pi f T_b} \right)^2. \tag{3.23}$$

Figure 3.15(b) plots the sinc^2 spectrum, revealing nulls at integer multiples of the bit rate, $1/T_b$, and "side lobes" beyond $f = \pm 1/T_b$.

What happens if this signal is applied to a low-pass filter having a narrow bandwidth, e.g., $1/(2T_b)$? Since the frequency components above $1/(2T_b)$ are suppressed, the signal experiences substantial ISI.

Let us continue our thought process and determine the spectrum if the binary sequence shown in Fig. 3.15(a) is impressed on the *phase* of a carrier. From the generation method of Fig. 3.13(b), we write

$$x_{PSK}(t) = x_{BB}(t)\cos\omega_c t, \tag{3.24}$$

concluding that the upconversion operation shifts the spectrum of $x_{BB}(t)$ to $\pm f_c = \pm\omega_c/(2\pi)$ (Fig. 3.16). From Fig. 3.13(a) and Example 3.1, we also recognize that the spectrum of an ASK waveform is similar but with impulses at $\pm f_c$.

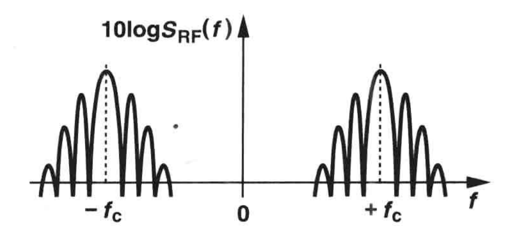


Figure 3.16 *Spectrum of PSK signal.*

Pulse Shaping The above analysis suggests that, to reduce the bandwidth of the modulated signal, the *baseband* pulse must be designed so as to occupy a small bandwidth itself. In this regard, the rectangular pulse used in the binary sequence of Fig. 3.15(a) is a poor choice: the sharp transitions between ZEROs and ONEs lead to an unnecessarily wide bandwidth. For this reason, the baseband pulses in communication systems are usually "shaped" to reduce their bandwidth. Shown in Fig. 3.17 is a conceptual example where the basic pulse exhibits smooth transitions, thereby occupying less bandwidth than rectangular pulses.

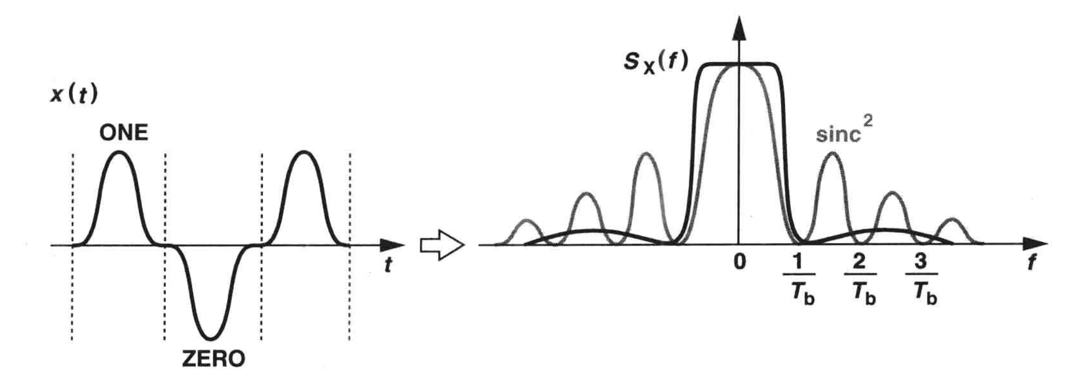


Figure 3.17 *Effect of smooth data transitions on spectrum.*

What pulse shape yields the tightest spectrum? Since the spectrum of an ideal rectangular pulse is a sinc, we surmise that a sinc pulse in the time domain gives a rectangular ("brickwall") spectrum [Fig. 3.18(a)]. Note that the spectrum is confined to $\pm 1/(2T_b)$. Now, if a random binary sequence employs such a pulse every T_b seconds, from Eq. (3.21) the spectrum still remains a rectangle [Fig. 3.18(b)] occupying substantially less bandwidth than $S_x(f)$ in Fig. 3.15(b). This bandwidth advantage persists after upconversion as well.

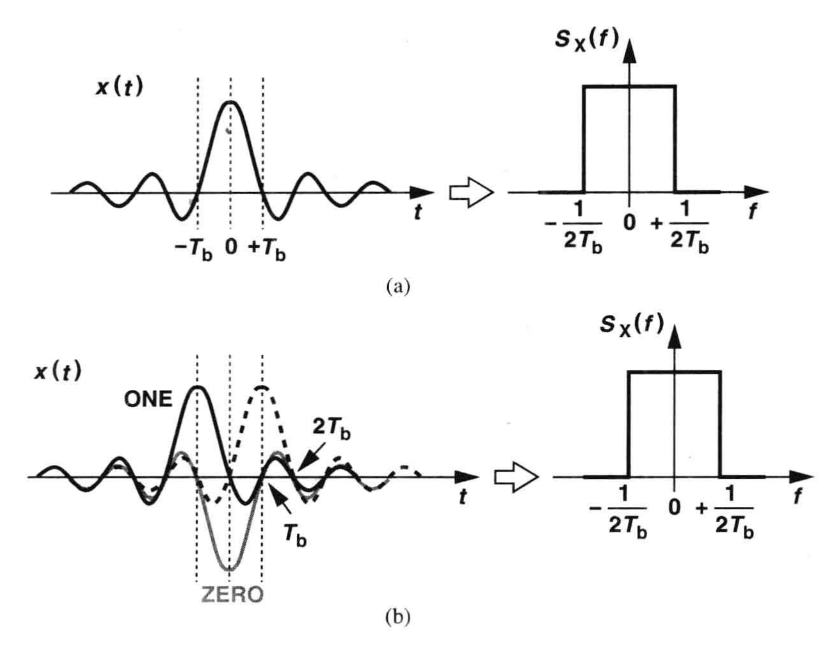


Figure 3.18 *(a) Sinc pulse and its spectrum, (b) random sequence of sinc pulses and its spectrum.*

Do we observe ISI in the random waveform of Fig. 3.18(b)? If the waveform is sampled at exactly integer multiples of T_b, then ISI is zero because all other pulses go through zero at these points. The use of such overlapping pulses that produce no ISI is called "Nyquist signaling." In practice, sinc pulses are difficult to generate and approximations are used instead. A common pulse shape is shown in Fig. 3.19(a) and expressed as

$$p(t) = \frac{\sin(\pi t/T_S)}{\pi t/T_S} \frac{\cos(\pi\alpha t/T_S)}{1 - 4\alpha^2 t^2/T_S^2}. \tag{3.25}$$

This pulse exhibits a "raised-cosine" spectrum [Fig. 3.19(b)]. Called the "roll-off factor," α determines how close $p(t)$ is to a sinc and, hence, the spectrum to a rectangle. For $\alpha = 0$, the pulse reduces to a sinc whereas for $\alpha = 1$, the spectrum becomes relatively wide. Typical values of α are in the range of 0.3 to 0.5.

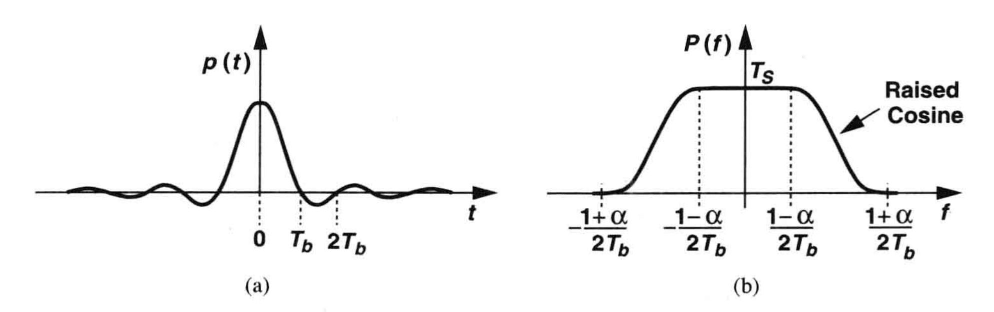


Figure 3.19 *Raised-cosine pulse shaping: (a) basic pulse and (b) corresponding spectrum.*

3.3.2 Signal Constellations

"Signal constellations" allow us to visualize modulation schemes and, more importantly, the effect of nonidealities on them. Let us begin with the binary PSK signal expressed by Eq. (3.24), which reduces to

$$x_{PSK}(t) = a_n \cos \omega_c t \qquad a_n = \pm 1 \tag{3.26}$$

for rectangular baseband pulses. We say this signal has one "basis function," $\cos \omega_c t$, and is simply defined by the possible values of the coefficient, a_n. Shown in Fig. 3.20(a), the constellation represents the values of a_n. The receiver must distinguish between these two values so as to decide whether the received bit is a ONE or a ZERO. In the presence of amplitude noise, the two points on the constellation become "fuzzy" as depicted in Fig. 3.20(b), sometimes coming closer to each other and making the detection more prone to error.



Figure 3.20 *Signal constellation for (a) ideal and (b) noisy PSK signal.*

Example 3.6

Plot the constellation of an ASK signal in the presence of amplitude noise.

Solution:

From the generation method of Fig. 3.13(a), we have

$$x_{ASK}(t) = a_n \cos \omega_c t \qquad a_n = 0,\ 1. \tag{3.27}$$

As shown in Fig. 3.21(a), noise corrupts the amplitude for both ZEROs and ONEs. Thus, the constellation appears as in Fig. 3.21(b).

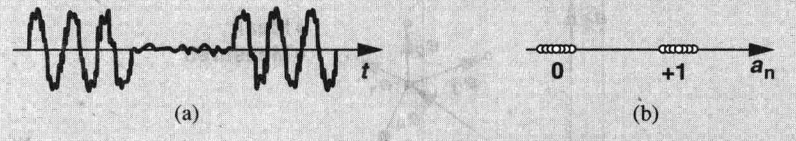


Figure 3.21 *(a) Noisy ASK signal and (b) its constellation.*

Next, we consider an FSK signal, which can be expressed as

$$x_{FSK}(t) = a_1 \cos \omega_1 t + a_2 \cos \omega_2 t \qquad a_1 a_2 = 10 \text{ or } 01. \tag{3.28}$$

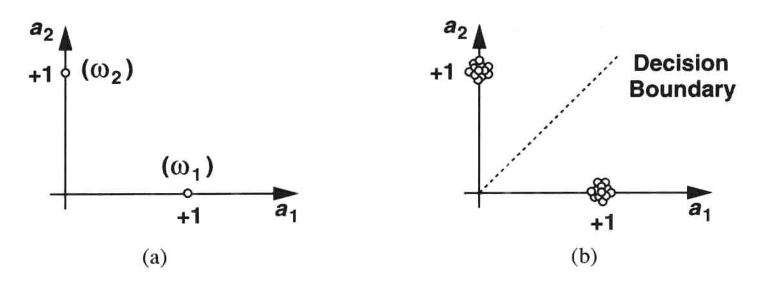


Figure 3.22 *Constellation of (a) ideal and (b) noisy FSK signal.*

We say $\cos \omega_1 t$ and $\cos \omega_2 t$ are the basis functions[4] and plot the possible values of a_1 and a_2 as in Fig. 3.22(a). An FSK receiver must decide whether the received frequency is ω_1 (i.e., $a_1 = 1$, $a_2 = 0$) or ω_2 (i.e., $a_1 = 0$, $a_2 = 1$). In the presence of noise, a "cloud" forms around each point in the constellation [Fig. 3.22(b)], causing an error if a particular sample crosses the decision boundary.

A comparison of the constellations in Figs. 3.20(b) and 3.22(b) suggests that PSK signals are less susceptible to noise than are FSK signals because their constellation points are farther from each other. This type of insight makes constellations a useful tool in analyzing RF systems.

The constellation can also provide a *quantitative* measure of the impairments that corrupt the signal. Representing the deviation of the constellation points from their ideal positions, the "error vector magnitude" (EVM) is such a measure. To obtain the EVM, a constellation based on a large number of detected samples is constructed and a vector is drawn between each measured point and its ideal position (Fig. 3.23). The EVM is defined as the rms magnitude of these error vectors normalized to the signal rms voltage:

$$\mathrm{EVM}_1 = \frac{1}{V_{rms}} \sqrt{\frac{1}{N} \sum_{j=1}^{N} e_j^2}, \tag{3.29}$$

where e_j denotes the magnitude of each error vector and V_{rms} the rms voltage of the signal. Alternatively, we can write

$$\mathrm{EVM}_2 = \frac{1}{P_{avg}} \cdot \frac{1}{N} \sum_{j=1}^{N} e_j^2, \tag{3.30}$$

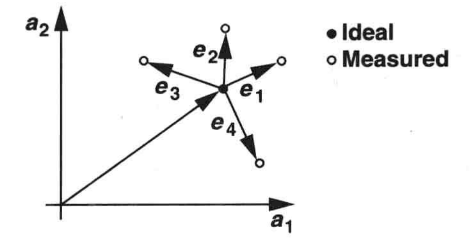


Figure 3.23 *Illustration of EVM.*

4. Basis functions must be orthogonal, i.e., have zero correlation.

where P_{avg} is the average signal power. Note that to express EVM in decibels, we compute $20 \log \text{EVM}_1$ or $10 \log \text{EVM}_2$.

The signal constellation and the EVM form a powerful tool for analyzing the effect of various nonidealities in the transceiver and the propagation channel. Effects such as noise, nonlinearity, and ISI readily manifest themselves in both.

3.3.3 Quadrature Modulation

Recall from Fig. 3.16 that binary PSK signals with square baseband pulses of width T_b seconds occupy a total bandwidth quite wider than $2/T_b$ hertz (after upconversion to RF). Baseband pulse shaping can decrease this bandwidth to about $2/T_b$.

In order to further reduce the bandwidth, "quadrature modulation," more specifically, "quadrature PSK" (QPSK) modulation can be performed. Illustrated in Fig. 3.24, the idea is to subdivide a binary data stream into pairs of two consecutive bits and impress these bits on the "quadrature phases" of the carrier, i.e., $\cos \omega_c t$ and $\sin \omega_c t$:

$$x(t) = b_{2m} A_c \cos \omega_c t - b_{2m+1} A_c \sin \omega_c t. \tag{3.31}$$

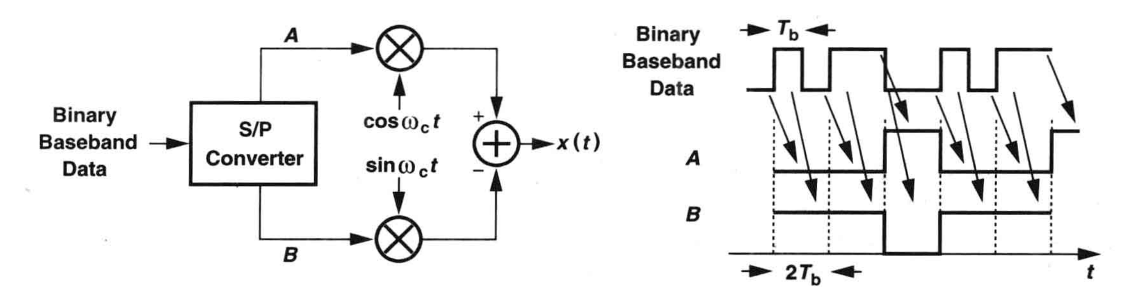


Figure 3.24 *Generation of QPSK signal.*

As shown in Fig. 3.24, a serial-to-parallel (S/P) converter (demultiplexer) separates the even-numbered bits, b_{2m}, and odd-numbered bits, b_{2m+1}, applying one group to the upper arm and the other to the lower arm. The two groups are then multiplied by the quadrature components of the carrier and subtracted at the output. Since $\cos \omega_c t$ and $\sin \omega_c t$ are orthogonal, the signal can be detected uniquely and the bits b_{2m} and b_{2m+1} can be separated without corrupting each other.

QPSK modulation halves the occupied bandwidth. This is simply because, as shown in Fig. 3.24, the demultiplexer "stretches" each bit duration by a factor of two before giving it to each arm. In other words, for a given pulse shape and bit rate, the spectra of PSK and QPSK are identical except for a bandwidth scaling by a factor of two. This is the principal reason for the widespread usage of QPSK. To avoid confusion, the pulses that appear at A and B in Fig. 3.24 are called "symbols" rather than bits.[5] Thus, the "symbol rate" of QPSK is half of its bit rate.

To obtain the QPSK constellation, we assume bits b_{2m} and b_{2m+1} are pulses with a height of ± 1 and write the modulated signal as $x(t) = \alpha_1 A_c \cos \omega_c t + \alpha_2 A_c \sin \omega_c t$, where

5. More precisely, the two consecutive bits that are demultiplexed and appear at A and B together form a symbol.

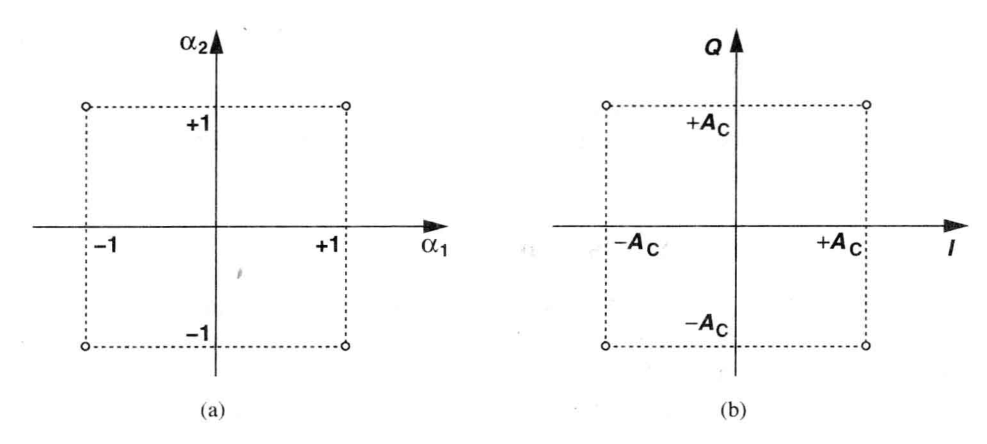


Figure 3.25 *QPSK signal constellation in terms of (a) α_1 and α_2, and (b) quadrature phases of carrier.*

α_1 and α_2 can each take on a value of $+1$ or -1. The constellation is shown in Fig. 3.25(a). More generally, the pulses appearing at A and B in Fig. 3.24 are called "quadrature baseband signals" and denoted by I (for "in-phase") and Q (for quadrature). For QPSK, $I = \alpha_1 A_c$ and $Q = \alpha_2 A_c$, yielding the constellation in Fig. 3.25(b). In this representation, too, we may simply plot the values of α_1 and α_2 in the constellation.

Example 3.7

Due to circuit nonidealities, one of the carrier phases in a QPSK modulator suffers from a small phase error ("mismatch") of θ:

$$x(t) = \alpha_1 A_c \cos(\omega_c t + \theta) + \alpha_2 A_c \sin \omega_c t. \tag{3.32}$$

Construct the signal constellation at the output of this modulator.

Solution:

We must reduce Eq. (3.32) to a form $\beta_1 A_c \cos \omega_c t + \beta_2 A_c \sin \omega_c t$:

$$x(t) = \alpha_1 A_c \cos\theta \cos \omega_c t + (\alpha_2 - \alpha_1 \sin\theta) A_c \sin \omega_c t. \tag{3.33}$$

Noting that α_1 and α_2 assume ± 1 values, we form four possible cases for the normalized coefficients of $\cos \omega_c t$ and $\sin \omega_c t$:

$$\beta_1 = +\cos\theta,\ \beta_2 = 1 - \sin\theta \tag{3.34}$$

$$\beta_1 = +\cos\theta,\ \beta_2 = -1 - \sin\theta \tag{3.35}$$

$$\beta_1 = -\cos\theta,\ \beta_2 = 1 + \sin\theta \tag{3.36}$$

$$\beta_1 = -\cos\theta,\ \beta_2 = -1 + \sin\theta. \tag{3.37}$$

Figure 3.26 superimposes the resulting constellation on the ideal one. As explained in Chapter 4, this distortion of the constellation becomes critical in both transmitters and receivers.

Example 3.7 (*Continued*)

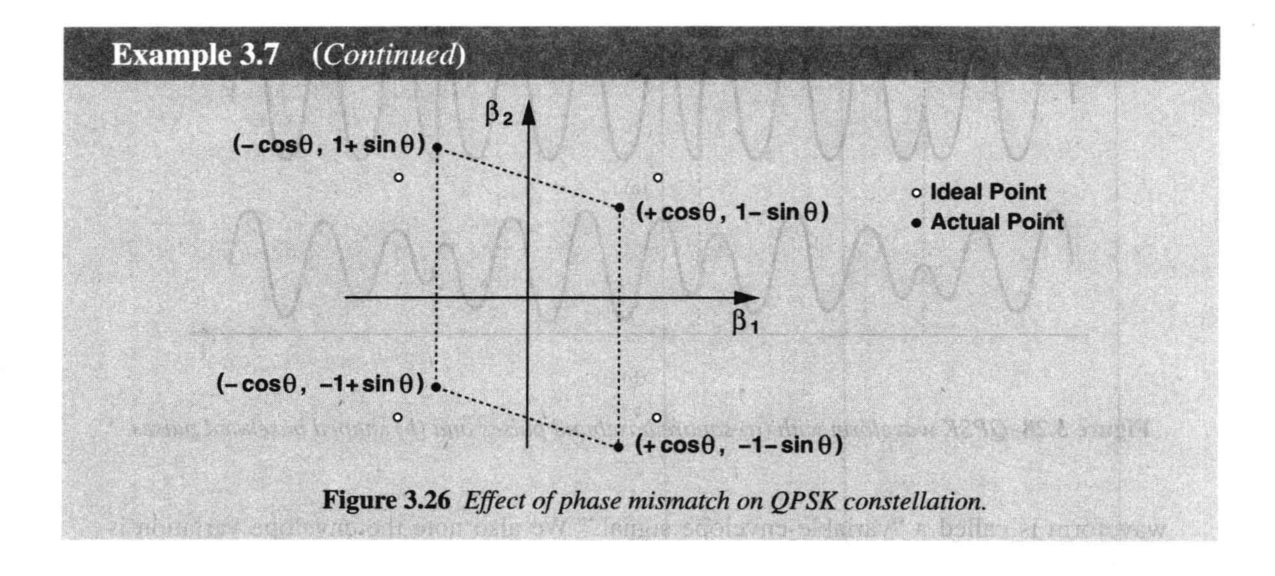


Figure 3.26 *Effect of phase mismatch on QPSK constellation.*

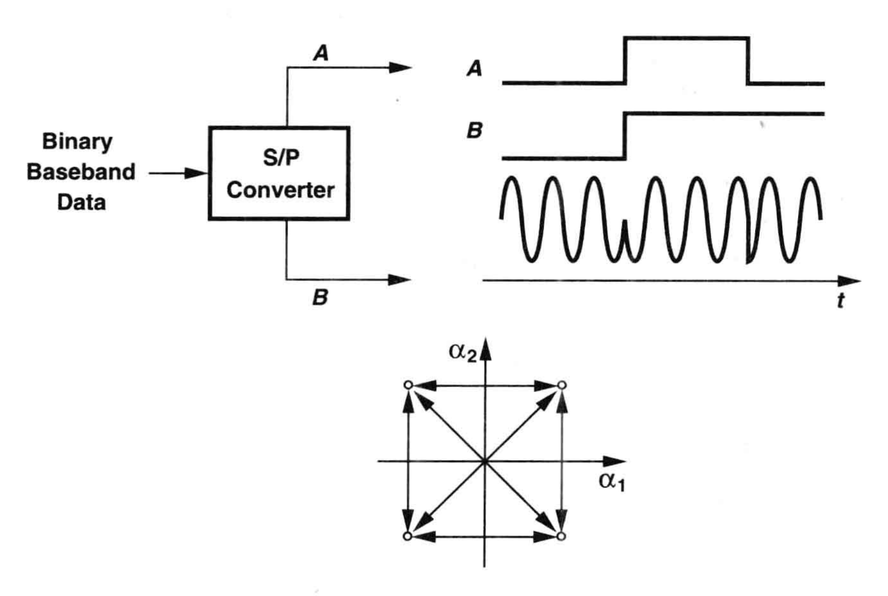


Figure 3.27 *Phase transitions in QPSK signal due to simultaneous transitions at A and B.*

An important drawback of QPSK stems from the large phase changes at the end of each symbol. As depicted in Fig. 3.27, when the waveforms at the output of the S/P converter change simultaneously from, say, $[-1 \quad -1]$ to $[+1 \quad +1]$, the carrier experiences a 180° phase step, or equivalently, a transition between two diagonally opposite points in the constellation. To understand why this is a serious issue, first recall from Section 3.3.1 that the baseband pulses are usually shaped so as to tighten the spectrum. What happens if the symbol pulses at nodes A and B are shaped before multiplication by the carrier phases? As illustrated in Fig. 3.28, with pulse shaping, the output signal amplitude ("envelope") experiences large changes each time the phase makes a 90° or 180° transition. The resulting

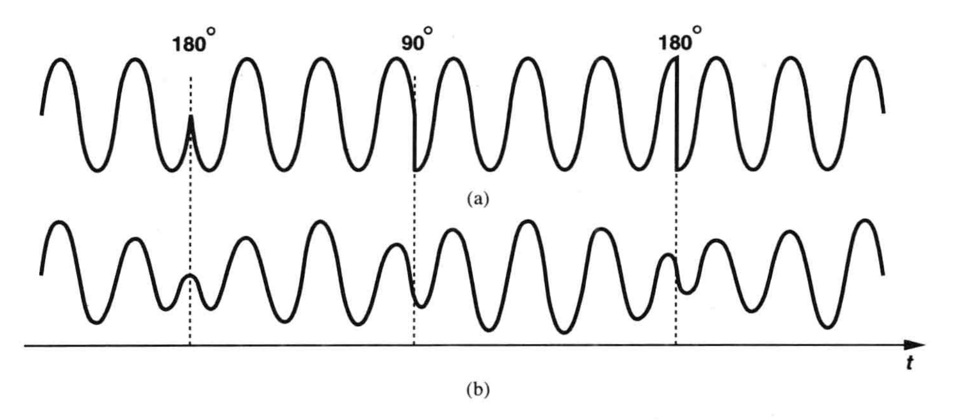


Figure 3.28 *QPSK waveform with (a) square baseband pulses and (b) shaped baseband pulses.*

waveform is called a "variable-envelope signal." We also note the envelope variation is proportional to the *phase change*. As explained in Section 3.4, a variable-envelope signal requires a *linear* power amplifier, which is inevitably less efficient than a nonlinear PA.

A variant of QPSK that remedies the above drawback is "offset QPSK" (OQPSK). As shown in Fig. 3.29, the data streams are offset in time by half the symbol period after S/P conversion, thereby avoiding simultaneous transitions in waveforms at nodes A and B. The phase step therefore does not exceed $\pm 90^\circ$. Figure 3.30 illustrates the phase transitions in the time domain and in the constellation. This advantage is obtained while maintaining the same spectrum. Unfortunately, however, OQPSK does not lend itself to "differential encoding" (Section 3.8). This type of encoding finds widespread usage as it obviates the need for "coherent detection," a difficult task.

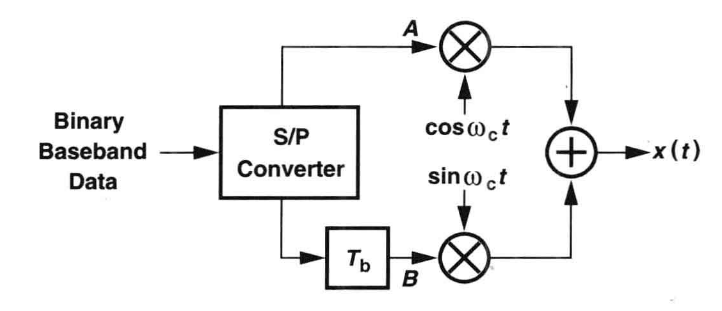


Figure 3.29 *Offset QPSK modulator.*

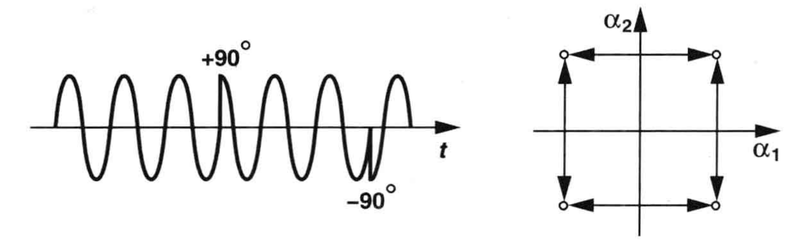


Figure 3.30 *Phase transitions in OQPSK.*

A variant of QPSK that can be differentially encoded is "$\pi/4$-QPSK" [4, 5]. In this case, the signal set consists of two QPSK schemes, one rotated 45° with respect to the other:

$$x_1(t) = A_c \cos\left(\omega_c t + k\frac{\pi}{4}\right) \quad k \text{ odd}, \tag{3.38}$$

$$x_2(t) = A_c \cos\left(\omega_c t + k\frac{\pi}{4}\right) \quad k \text{ even}. \tag{3.39}$$

As shown in Fig. 3.31, the modulation is performed by alternately taking the output from each QPSK generator.

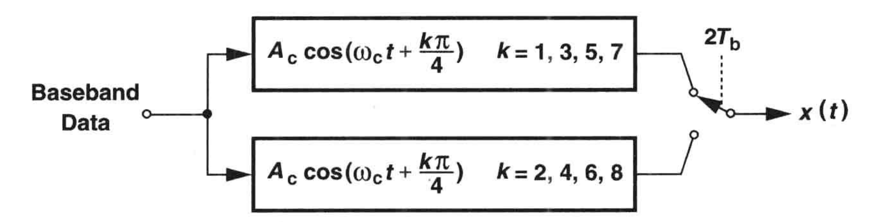


Figure 3.31 *Conceptual generation of $\pi/4$-QPSK signal.*

To better understand the operation, let us study the simple $\pi/4$-QPSK generator shown in Fig. 3.32. After S/P conversion, the digital signal levels are scaled and shifted so as to present ± 1 in the upper QPSK modulator and 0 and $\sqrt{2}$ in the lower. The outputs are therefore equal to $x_1(t) = \alpha_1 \cos\omega_c t + \alpha_2 \sin\omega_c t$, where $[\alpha_1\ \alpha_2] = [\pm A_c\ \pm A_c]$, and $x_2(t) = \beta_1 \cos\omega_c t + \beta_2 \sin\omega_c t$, where $[\beta_1\ \beta_2] = [0\ \pm\sqrt{2}A_c]$ and $[\pm\sqrt{2}A_c\ 0]$. Thus, the constellation alternates between the two depicted in Fig. 3.32. Now consider a baseband sequence of [11, 01, 10, 11, 01]. As shown in Fig. 3.33, the first pair, [1 1], is converted to $[+A_c\ +A_c]$ in the upper arm, producing $y(t) = A_c \cos(\omega_c t + \pi/4)$. The next pair, [0 1], is converted to $[0\ -\sqrt{2}A_c]$ in the lower arm, yielding $y(t) = -\sqrt{2}A_c \cos\omega_c t$. Following

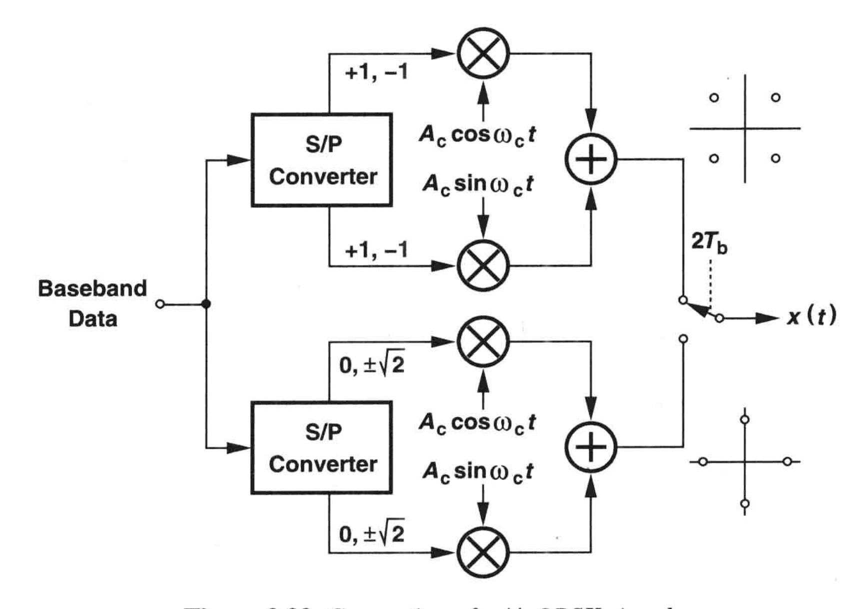


Figure 3.32 *Generation of $\pi/4$-QPSK signals.*

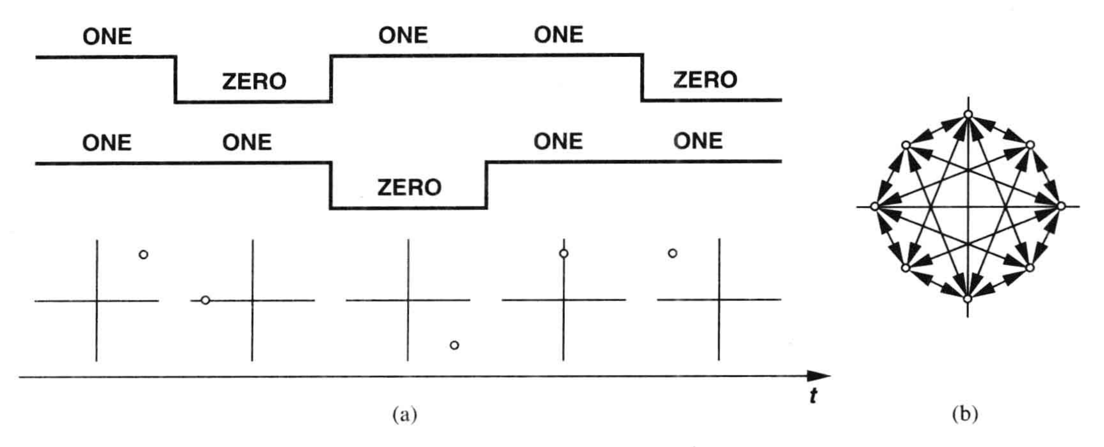


Figure 3.33 *(a) Evolution of $\pi/4$-QPSK in time domain, (b) possible phase transitions in the constellation.*

the values of $y(t)$ for the entire sequence, we note that the points chosen from the two constellations appear as in Fig. 3.33(a) as a function of time. The key point here is that, since no two consecutive points are from the same constellation, the maximum phase step is 135°, 45° less than that in QPSK. This is illustrated in Fig. 3.33(b). Thus, in terms of the maximum phase change, $\pi/4$-QPSK is an intermediate case between QPSK and OQPSK.

By virtue of baseband pulse shaping, QPSK and its variants provide high spectral efficiency but lead to poor power efficiency because they dictate linear power amplifiers. These modulation schemes are used in a number of applications (Section 3.7).

3.3.4 GMSK and GFSK Modulation

A class of modulation schemes that does not require linear power amplifiers, thus exhibiting high power efficiency, is "constant-envelope modulation." For example, an FSK waveform expressed as $x_{FSK}(t) = A_c \cos[\omega_c t + m \int x_{BB}(t)dt]$ has a constant envelope. To arrive at variants of FSK, let us first consider the implementation of a frequency modulator. As illustrated in Fig. 3.34(a), an oscillator whose frequency can be tuned by a voltage [called a "voltage-controlled oscillator" (VCO)] performs frequency modulation. In FSK, square baseband pulses are applied to the VCO, producing a broad output spectrum due to the abrupt changes in the VCO frequency. We therefore surmise that smoother transitions between ONEs and ZEROs in the baseband signal can tighten the spectrum. A common method of pulse shaping for frequency modulation employs a "Gaussian filter," i.e., one whose impulse response is a Gaussian pulse. Thus, as shown in Fig. 3.34(b), the pulses applied to the VCO gradually change the output frequency, leading to a narrower spectrum.

Called "Gaussian minimum shift keying" (GMSK), the scheme of Fig. 3.34(b) is used in GSM cell phones (Section 3.7). The GMSK waveform can be expressed as

$$x_{GMSK}(t) = A_c \cos[\omega_c t + m \int x_{BB}(t) * h(t)dt], \qquad (3.40)$$

where $h(t)$ denotes the impulse response of the Gaussian filter. The modulation index, m, is a dimensionless quantity and has a value of 0.5. Owing to its constant envelope, GMSK allows optimization of PAs for high efficiency—with little attention to linearity.

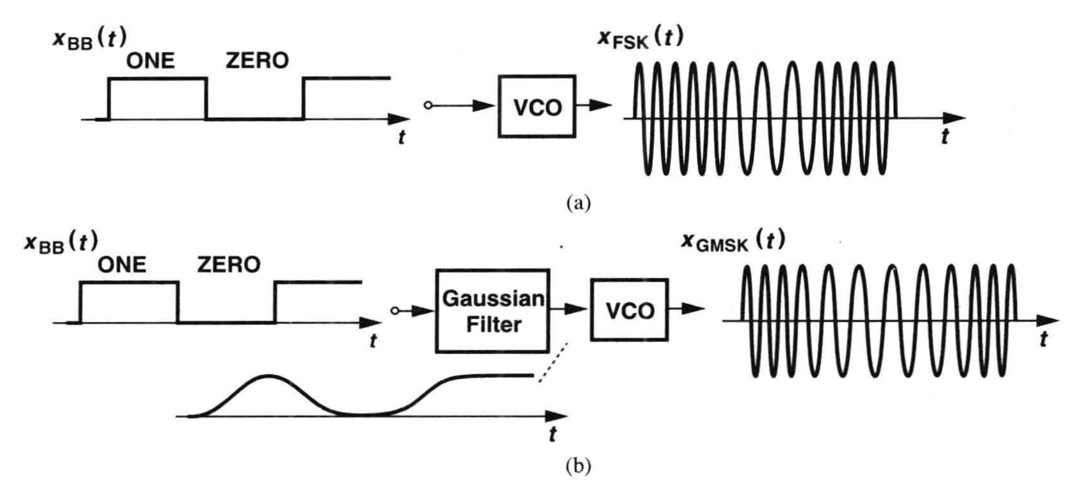


Figure 3.34 *Generation of (a) FSK and (b) GMSK signals.*

A slightly different version of GMSK, called Gaussian frequency shift keying (GFSK), is employed in Bluetooth. The GFSK waveform is also given by Eq. (3.40) but with $m = 0.3$.

Example 3.8

Construct a GMSK modulator using a quadrature upconverter.

Solution:

Let us rewrite Eq. (3.40) as

$$x_{GMSK}(t) = A_c \cos[m \int x_{BB}(t) * h(t)dt] \cos \omega_c t - A_c \sin[m \int x_{BB}(t) * h(t)dt] \sin \omega_c t. \quad (3.41)$$

We can therefore construct the modulator as shown in Fig. 3.35, where a Gaussian filter is followed by an integrator and two arms that compute the sine and cosine of the signal at node A. The complexity of these operations is much more easily afforded in the digital domain than in the analog domain (Chapter 4).

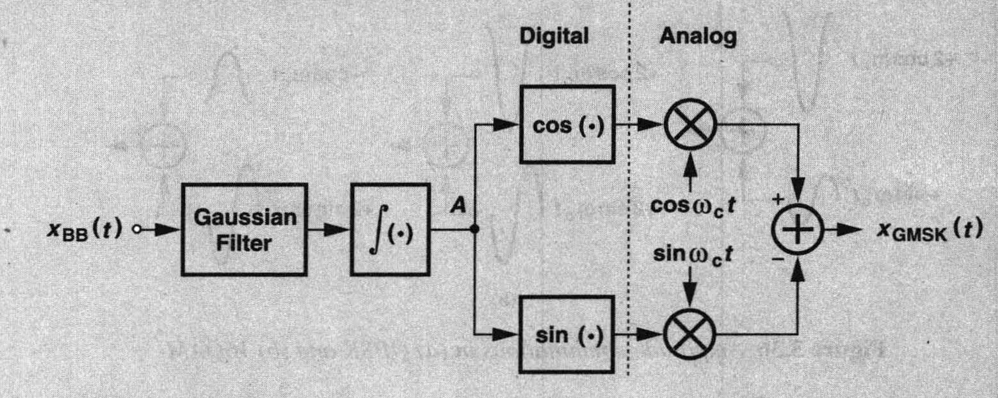


Figure 3.35 *Mixed-mode generation of GMSK signal.*

3.3.5 Quadrature Amplitude Modulation

Our study of PSK and QPSK has revealed a twofold reduction in the spectrum as a result of impressing the information on the quadrature components of the carrier. Can we extend this idea to further tighten the spectrum? A method that accomplishes this goal is called "quadrature amplitude modulation" (QAM).

To arrive at QAM, let us first draw the four possible waveforms for QPSK corresponding to the four points in the constellation. As predicted by Eq. (3.31) and shown in Fig. 3.36(a), each quadrature component of the carrier is multipled by $+1$ or -1 according to the values of b_{2m} and b_{2m+1}. Now suppose we allow *four* possible amplitudes for the sine and cosine waveforms, e.g., ± 1 and ± 2, thus obtaining 16 possible output waveforms. Figure 3.36(b) depicts a few examples of such waveforms. In other words, we group *four* consecutive bits of the binary baseband stream and select one of the 16 waveforms accordingly. Called "16QAM,"[6] the resulting output occupies *one-fourth* the bandwidth of PSK and is expressed as

$$x_{16QAM}(t) = \alpha_1 A_c \cos \omega_c t - \alpha_2 A_c \sin \omega_c t \quad \alpha_1 = \pm 1, \pm 2, \ \alpha_2 = \pm 1, \pm 2. \tag{3.42}$$

The constellation of 16QAM can be constructed using the 16 possible combinations of $[\alpha_1 \ \alpha_2]$ (Fig. 3.37). For a given transmitted power [e.g., the rms value of the waveforms shown in Fig. 3.36(b)], the points in this constellation are closer to one another than those in the QPSK constellation, making the detection more sensitive to noise. This is the price paid for saving bandwidth.

In addition to a "dense" constellation, 16QAM also exhibits large envelope variations, as exemplified by the waveforms in Fig. 3.36. Thus, this type of modulation requires a

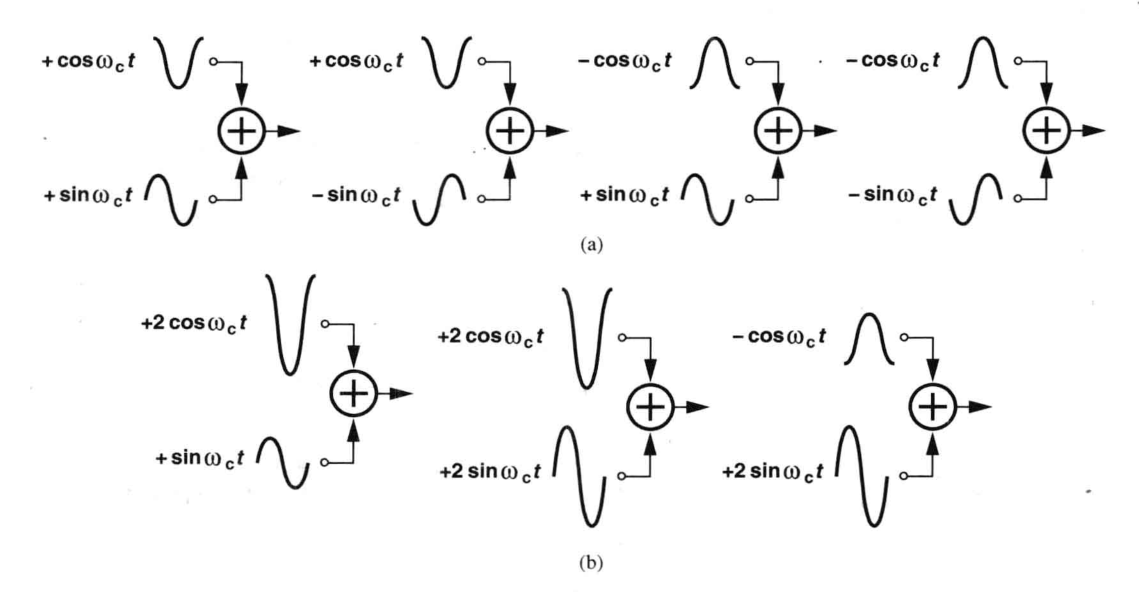


Figure 3.36 *Amplitude combinations in (a) QPSK and (b) 16QAM.*

6. Also known as QAM16.

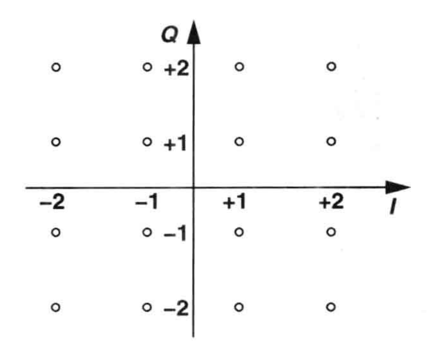


Figure 3.37 *Constellation of 16QAM signal.*

highly-linear power amplifier. We again observe the trade-offs among bandwidth efficiency, detectability, and power efficiency.

The concept of QAM can be extended to even denser constellations. For example, if *eight* consecutive bits in the binary baseband stream are grouped and, accordingly, each quadrature component of the carrier is allowed to have eight possible amplitudes, then 64QAM is obtained. The bandwidth is therefore reduced by a factor of eight with respect to that of PSK, but the detection and power amplifier design become more difficult.

A number of applications employ QAM to save bandwidth. For example, IEEE802.11g/a uses 64QAM for the highest data rate (54 Mb/s).

3.3.6 Orthogonal Frequency Division Multiplexing

Communication in a wireless environment entails a serious issue called "multipath propagation." Illustrated in Fig. 3.38(a), this effect arises from the propagation of the electromagnetic waves from the transmitter to the receiver through *multiple paths*. For example, one wave directly propagates from the TX to the RX while another is reflected from a wall before reaching the receiver. Since the phase shift associated with reflection(s) depends on both the path length and the reflecting material, the waves arrive at the RX with vastly different *delays*, or a large "delay spread." Even if these delays do not result in destructive interference of the rays, they may lead to considerable *intersymbol* interference. To understand this point, suppose, for example, two ASK waveforms containing the same information reach the RX with different delays [Fig. 3.38(b)]. Since the antenna senses the *sum* of these waveforms, the baseband data consists of two copies of the signal that are shifted in time, thus experiencing ISI [Fig. 3.38(c)].

The ISI resulting from multipath effects worsens for larger delay spreads or higher bit rates. For example, a data rate of 1 Mb/s becomes sensitive to multipath propagation if the delay spread reaches a fraction of a microsecond. As a rule of thumb, we say communication inside office buildings and homes begins to suffer from multipath effects for data rates greater than 10 Mb/s.

How does wireless communication handle higher data rates? An interesting method of delay spread mitigation is called "orthogonal frequency division multiplexing" (OFDM). Consider the "single-carrier" modulated spectrum shown in Fig. 3.39(a), which occupies a relatively large bandwidth due to a high data rate of r_b bits per second. In OFDM, the baseband data is first demultiplexed by a factor of N, producing N streams each having

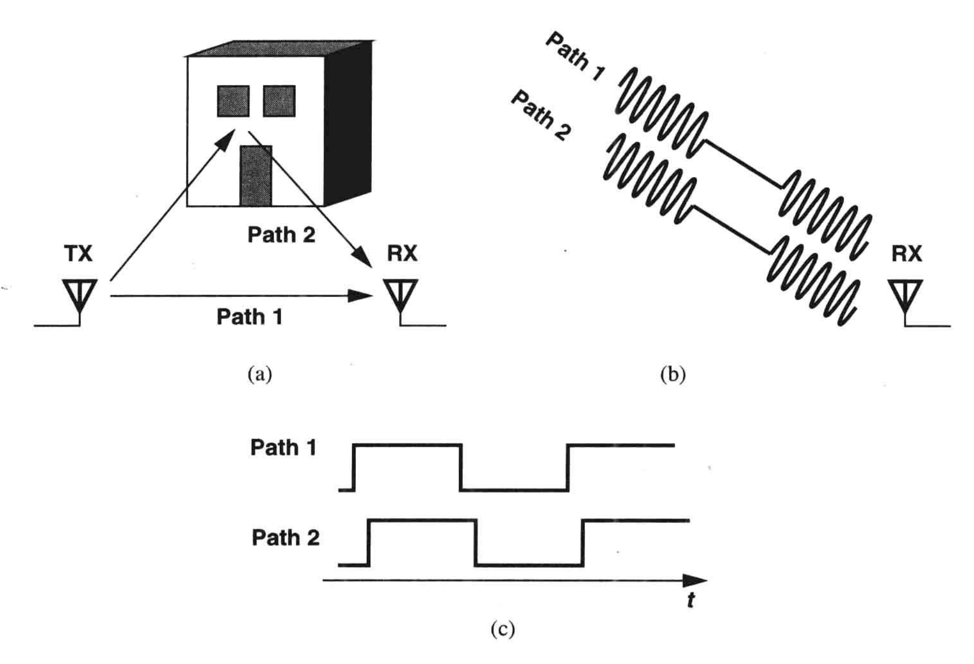


Figure 3.38 *(a) Multipath propagation, (b) effect on received ASK waveforms, (c) baseband components exhibiting ISI due to delay spread.*

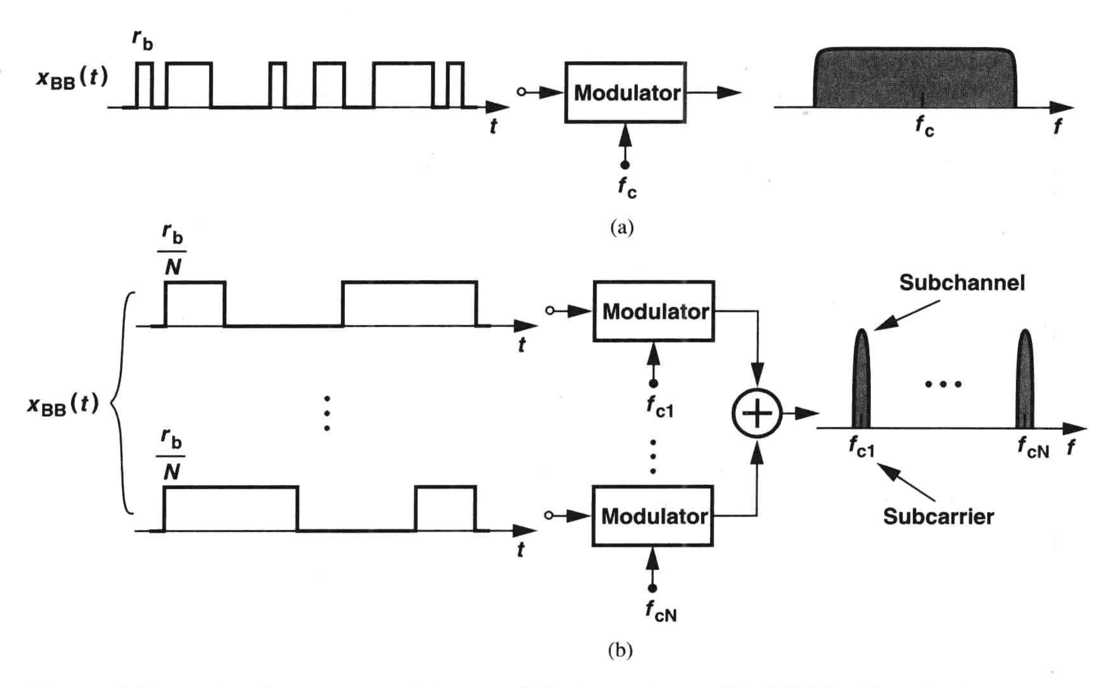


Figure 3.39 *(a) Single-carrier modulator with high-rate input, (b) OFDM with multiple carriers.*

a (symbol) rate of r_b/N [Fig. 3.39(b)]. The N streams are then impressed on N *different* carrier frequencies, f_{c1}-f_{cN}, leading to a "multi-carrier" spectrum. Note that the total bandwidth and data rate remain equal to those of the single-carrier spectrum, but the multi-carrier signal is less sensitive to multipath effects because each carrier contains a low-rate data stream and can therefore tolerate a larger delay spread.

Each of the N carriers in Fig. 3.39(b) is called a "subcarrier" and each resulting modulated output a "subchannel." In practice, all of the subchannels utilize the same modulation scheme. For example, IEEE802.11a/g employs 48 subchannels with 64QAM in each for the highest data rate (54 Mb/s). Thus, each subchannel carries a symbol rate of (54 Mb/s)/48/8=141 ksymbol/s.

Example 3.9

It appears that an OFDM transmitter is very complex as it requires tens of carrier frequencies and modulators (i.e., tens of oscillators and mixers). How is OFDM realized in practice?

Solution:

In practice, the subchannel modulations are performed in the digital baseband and subsequently converted to analog form. In other words, rather than generate $a_1(t)\cos[\omega_c t + \phi_1(t)] + a_2(t)\cos[\omega_c t + \Delta\omega t + \phi_2(t)] + \cdots$, we first construct $a_1(t)\cos\phi_1(t) + a_2(t)\cos[\Delta\omega t + \phi_2(t)] + \cdots$ and $a_1(t)\sin\phi_1(t) + a_2(t)\sin[\Delta\omega t + \phi_2(t)] + \cdots$. These components are then applied to a quadrature modulator with an LO frequency of ω_c.

While providing greater immunity to multipath propagation, OFDM imposes severe linearity requirements on power amplifiers. This is because the N (orthogonal) subchannels summed at the output of the system in Fig. 3.39(b) may add constructively at some point in time, creating a large amplitude, and destructively at some other point in time, producing a small amplitude. That is, OFDM exhibits large envelope variations even if the modulated waveform in each subchannel does not.

In the design of power amplifiers, it is useful to have a quantitative measure of the signal's envelope variations. One such measure is the "peak-to-average ratio" (PAR). As illustrated in Fig. 3.40, PAR is defined as the ratio of the largest value of the square of the

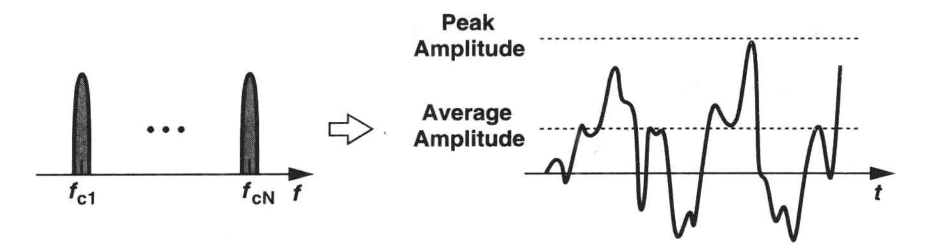


Figure 3.40 *Large amplitude variations due to OFDM.*

signal (voltage or current) divided by the average value of the square of the signal:

$$\text{PAR} = \frac{\text{Max}[x^2(t)]}{\overline{x^2(t)}}. \tag{3.43}$$

We note that three effects lead to a large PAR: pulse shaping in the baseband, amplitude modulation schemes such as QAM, and orthogonal frequency division multiplexing. For N subcarriers, the PAR of an OFDM waveform is about $2 \ln N$ if N is large [6].

3.4 SPECTRAL REGROWTH

In our study of modulation schemes, we have mentioned that variable-envelope signals require linear PAs, whereas constant-envelope signals do not. Of course, modulation schemes such as 16QAM that carry information in their amplitude levels experience corruption if the PA compresses the larger levels, i.e., moves the outer points of the constellation toward the origin. But even variable-envelope signals that carry no significant information in their amplitude (e.g., QPSK with baseband pulse-shaping) create an undesirable effect in nonlinear PAs. Called "spectral regrowth," this effect corrupts the adjacent channels.

A modulated waveform $x(t) = A(t)\cos[\omega_c t + \phi(t)]$ is said to have a constant envelope if $A(t)$ does not vary with time. Otherwise, we say the signal has a variable envelope. Constant- and variable-envelope signals behave differently in a nonlinear system. Suppose $A(t) = A_c$ and the system exhibits a third-order memoryless nonlinearity:

$$y(t) = \alpha_3 x^3(t) + \cdots \tag{3.44}$$

$$= \alpha_3 A_c^3 \cos^3[\omega_c t + \phi(t)] + \cdots \tag{3.45}$$

$$= \frac{\alpha_3 A_c^3}{4}\cos[3\omega_c t + 3\phi(t)] + \frac{3\alpha_3 A_c^3}{4}\cos[\omega_c t + \phi(t)] \tag{3.46}$$

The first term in (3.46) represents a modulated signal around $\omega = 3\omega_c$. Since the bandwidth of the original signal, $A_c \cos[\omega_c t + \phi(t)]$, is typically much less than ω_c, the bandwidth occupied by $\cos[3\omega_c t + 3\phi(t)]$ is small enough that it does not reach the center frequency of ω_c. Thus, the shape of the spectrum in the vicinity of ω_c remains unchanged.

Now consider a variable-envelope signal applied to the above nonlinear system. Writing $x(t)$ as

$$x(t) = x_I(t)\cos\omega_c t - x_Q(t)\sin\omega_c t, \tag{3.47}$$

where x_I and $x_Q(t)$ are the baseband I and Q components, we have

$$y(t) = \alpha_3[x_I(t)\cos\omega_c t - x_Q(t)\sin\omega_c t]^3 + \cdots \tag{3.48}$$

$$= \alpha_3 x_I^3(t)\frac{\cos 3\omega_c t + 3\cos\omega_c t}{4} - \alpha_3 x_Q^3(t)\frac{-\cos 3\omega_c t + 3\sin\omega_c t}{4}. \tag{3.49}$$

Thus, the output contains the spectra of $x_I^3(t)$ and $x_Q^3(t)$ centered around ω_c. Since these components generally exhibit a broader spectrum than do $x_I(t)$ and $x_Q(t)$, we say the spectrum "grows" when a variable-envelope signal passes through a nonlinear system. Figure 3.41 summarizes our findings.

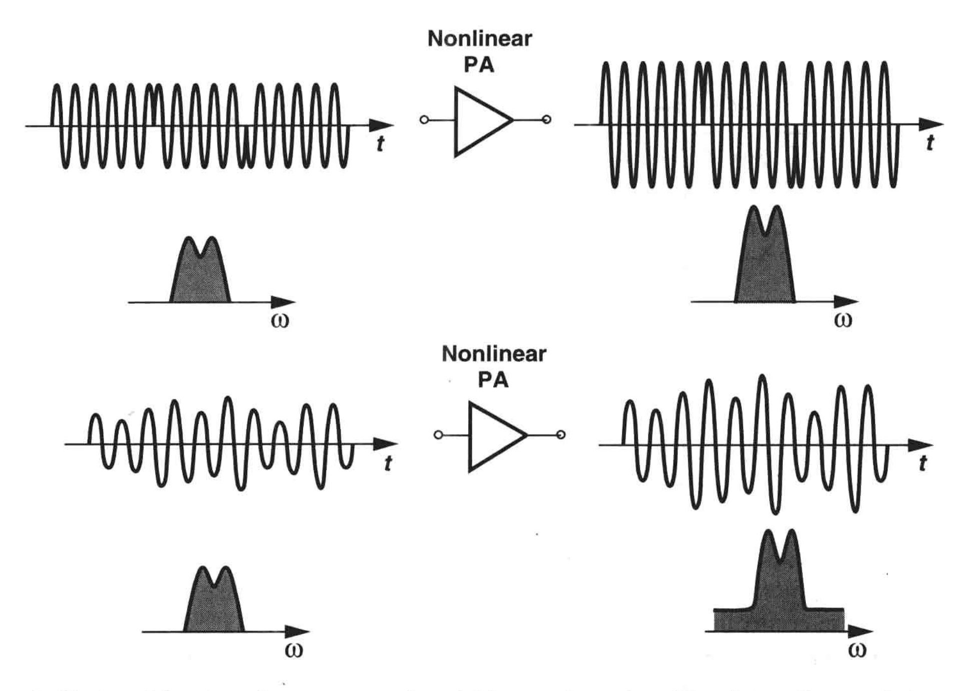


Figure 3.41 *Amplification of constant- and variable-envelope signals and the effect on their spectra.*

3.5 MOBILE RF COMMUNICATIONS

A mobile system is one in which users can physically move while communicating with one another. Examples include pagers, cellular phones, and cordless phones. It is the mobility that has made RF communications powerful and popular. The transceiver carried by the user is called the "mobile unit" (or simply the "mobile"), the "terminal," or the "hand-held unit." The complexity of the wireless infrastructure often demands that the mobiles communicate only through a fixed, relatively expensive unit called the "base station." Each mobile receives and transmits information from and to the base station via two RF channels called the "forward channel" or "downlink" and the "reverse channel" or "uplink," respectively. Most of our treatment in this book relates to the mobile unit because, compared to the base station, hand-held units constitute a much larger portion of the market and their design is much more similar to other types of RF systems.

Cellular System With the limited available spectrum (e.g., 25 MHz around 900 MHz), how do hundreds of thousands of people communicate in a crowded metropolitan area? To answer this question, we first consider a simpler case: thousands of FM radio broadcasting stations may operate in a country in the 88–108 MHz band. This is possible because stations that are physically far enough from each other can use the same carrier frequency ("frequency reuse") with negligible mutual interference (except at some point in the middle where the stations are received with comparable signal levels). The minimum distance between two stations that can employ equal carrier frequencies depends on the signal power produced by each.

In mobile communications, the concept of frequency reuse is implemented in a "cellular" structure, where each cell is configured as a hexagon and surrounded by 6 other cells

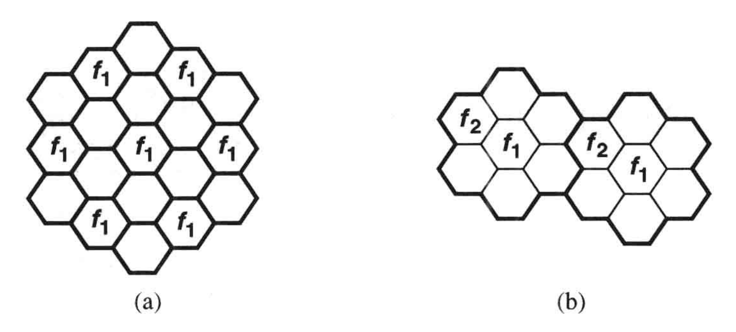


Figure 3.42 *(a) Simple cellular system, (b) 7-cell reuse pattern.*

[Fig. 3.42(a)]. The idea is that if the center cell uses a frequency f_1 for communication, the 6 neighboring cells cannot utilize this frequency, but the cells beyond the immediate neighbors may. In practice, more efficient frequency assignment leads to the "7-cell" reuse pattern shown in Fig. 3.42(b). Note that in reality each cell utilizes a group of frequencies.

The mobile units in each cell of Fig. 3.42(b) are served by a base station, and all of the base stations are controlled by a "mobile telephone switching office" (MTSO).

Co-Channel Interference An important issue in a cellular system is how much two cells that use the same frequency interfere with each other (Fig. 3.43). Called "co-channel interference" (CCI), this effect depends on the *ratio* of the distance between two co-channel cells to the cell radius and is independent of the transmitted power. Given by the frequency reuse plan, this ratio is approximately equal to 4.6 for the 7-cell pattern of Fig. 3.42(b) [7]. It can be shown that this value yields a signal-to-co-channel interference ratio of 18 dB [7].

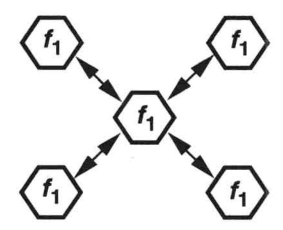


Figure 3.43 *Co-channel interference.*

Hand-off What happens when a mobile unit "roams" from cell A to cell B (Fig. 3.44)? Since the power level received from the base station in cell A is insufficient

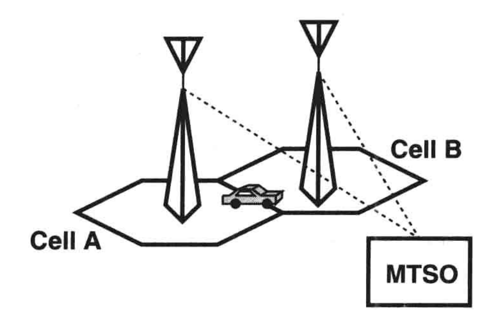


Figure 3.44 *Problem of hand-off.*

to maintain communication, the mobile must change its server to the base station in cell B. Furthermore, since adjacent cells do not use the same group of frequencies, the channel must also change. Called "hand-off," this process is performed by the MTSO. Once the level received by the base station in cell A drops below a threshold, the MTSO hands off the mobile to the base station in cell B, hoping that the latter is close enough. This strategy fails with relatively high probability, resulting in dropped calls.

To improve the hand-off process, second-generation cellular systems allow the mobile unit to measure the received signal level from different base stations, thus performing hand-off when the path to the second base station has sufficiently low loss [7].

Path Loss and Multipath Fading Propagation of signals in a mobile communication environment is quite complex. We briefly describe some of the important concepts here. Signals propagating through free space experience a power loss proportional to the square of the distance, d, from the source. In reality, however, the signal travels through both a direct path and an indirect, reflective path (Fig. 3.45). It can be shown that in this case, the loss increases with the *fourth* power of the distance [7]. In crowded areas, the actual loss profile may be proportional to d^2 for some distance and d^4 for another.

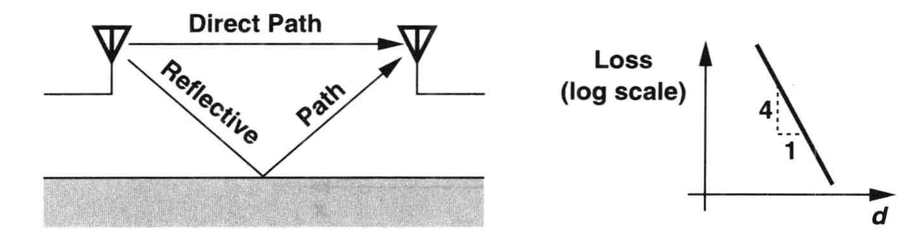


Figure 3.45 *Indirect signal propagation and resulting loss profile.*

In addition to the overall loss profile depicted in Fig. 3.45, another mechanism gives rise to fluctuations in the received signal level as a function of distance. Since the two signals shown in Fig. 3.45 generally experience different phase shifts, possibly arriving at the receiver with opposite phases and roughly equal amplitudes, the net received signal may be very small. Called "multipath fading" and illustrated in Fig. 3.46, this phenomenon introduces enormous variations in the signal level as the receiver moves by a fraction of the wavelength. Note that multipath propagation creates fading and/or ISI.

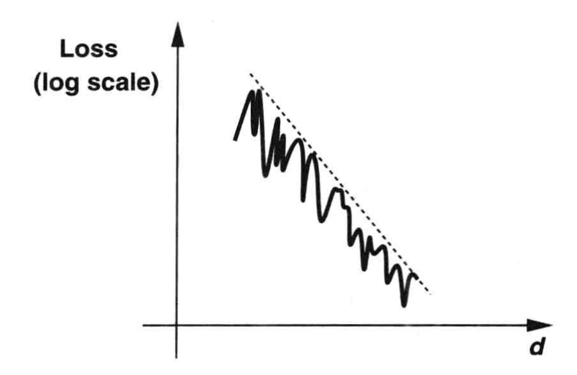


Figure 3.46 *Multipath loss profile.*

In reality, since the transmitted signal is reflected by many buildings and moving cars, the fluctuations are quite irregular. Nonetheless, the overall received signal can be expressed as

$$x_R(t) = a_1(t)\cos(\omega_c t + \theta_1) + a_2(t)\cos(\omega_c t + \theta_2) + \cdots + a_n\cos(\omega_c t + \theta_n) \quad (3.50)$$

$$= \left[\sum_{j=1}^{n} a_j(t)\cos\theta_j\right]\cos\omega_c t - \left[\sum_{j=1}^{n} a_j(t)\sin\theta_j\right]\sin\omega_c t. \quad (3.51)$$

For large n, each summation has a Gaussian distribution. Denoting the first summation by A and the second by B, we have

$$x_R(t) = \sqrt{A^2 + B^2}\cos(\omega_c t + \phi), \quad (3.52)$$

where $\phi = \tan^{-1}(B/A)$. It can be shown that the amplitude, $A_m = \sqrt{A^2 + B^2}$, has a Rayleigh distribution (Fig. 3.47) [1], exhibiting losses greater than 10 dB below the mean for approximately 6% of the time.

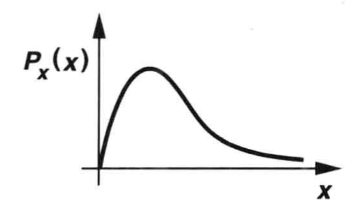


Figure 3.47 *Rayleigh distribution.*

From the above discussion, we conclude that in an RF system the transmitter output power and the receiver dynamic range must be chosen so as to accommodate signal level variations due to both the overall path loss (roughly proportional to d^4) and the deep multipath fading effects. While it is theoretically possible that multipath fading yields a zero amplitude (infinite loss) at a given distance, the probability of this event is negligible because moving objects in a mobile environment tend to "soften" the fading.

Diversity The effect of fading can be lowered by adding redundancy to the transmission or reception of the signal. "Space diversity" or "antenna diversity" employs two or more antennas spaced apart by a significant fraction of the wavelength so as to achieve a higher probability of receiving a nonfaded signal.

"Frequency diversity" refers to the case where multiple carrier frequencies are used, with the idea that fading is unlikely to occur simultaneously at two frequencies sufficiently far from each other. "Time diversity" is another technique whereby the data is transmitted or received more than once to overcome short-term fading.

Delay Spread Suppose two signals in a multipath environment experience roughly equal attenuations but different delays. This is possible because the absorption coefficient and phase shift of reflective or refractive materials vary widely, making it likely for two paths to exhibit equal loss and unequal delays. Addition of two such signals yields $x(t) = A\cos\omega(t-\tau_1) + A\cos\omega(t-\tau_2) = 2A\cos[(2\omega t - \omega\tau_1 - \omega\tau_2)/2]\cos[\omega(\tau_1 - \tau_2)/2]$, where the second cosine factor relates the fading to the "delay spread," $\Delta\tau = \tau_1 - \tau_2$. An

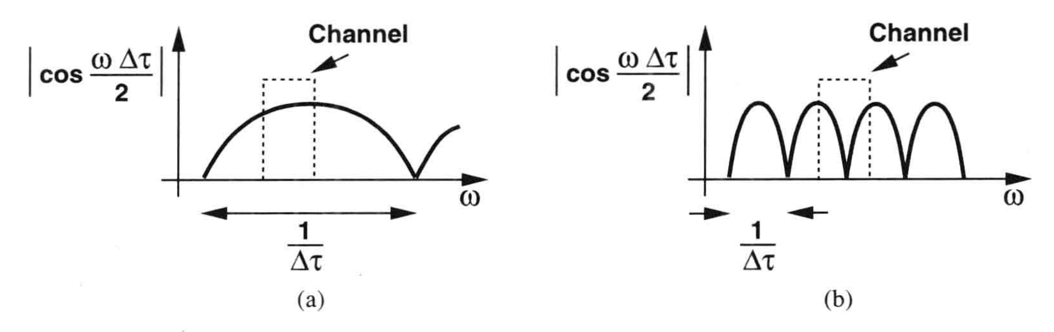


Figure 3.48 *(a) Flat and (b) frequency-selective fading.*

important issue here is the frequency dependence of the fade. As illustrated in Fig. 3.48, small delay spreads yield a relatively flat fade, whereas large delay spreads introduce considerable variation in the spectrum.

In a multipath environment, many signals arrive at the receiver with different delays, yielding rms delay spreads as large as several microseconds and hence fading bandwidths of several hundreds of kilohertz. Thus, an entire communication channel may be suppressed during such a fade.

Interleaving The nature of multipath fading and the signal processing techniques used to alleviate this issue is such that errors occur in clusters of bits. In order to lower the effect of these errors, the baseband bit stream in the transmitter undergoes "interleaving" before modulation. An interleaver in essence scrambles the time order of the bits according to an algorithm known by the receiver [7].

3.6 MULTIPLE ACCESS TECHNIQUES

3.6.1 Time and Frequency Division Duplexing

The simplest case of multiple access is the problem of two-way communication by a transceiver, a function called "duplexing." In old walkie-talkies, for example, the user would press the "talk" button to transmit while disabling the receive path and release the button to listen while disabling the transmit path. This can be considered a simple form of "time division duplexing," (TDD), whereby the same frequency band is utilized for both transmit (TX) and receive (RX) paths but the system transmits for half of the time and receives for the other half. Illustrated in Fig. 3.49, TDD is usually performed fast enough to be transparent to the user.

Another approach to duplexing is to employ two different frequency bands for the transmit and receive paths. Called "frequency-division duplexing" (FDD) and shown in Fig. 3.50, this technique incorporates band-pass filters to isolate the two paths, allowing simultaneous transmission and reception. Since two such transceivers cannot communicate directly, the TX band must be translated to the RX band at some point. In wireless networks, this translation is performed in the base station.

It is instructive to contrast the two duplexing methods by considering their merits and drawbacks. In TDD, an RF switch with a loss less than 1 dB follows the antenna to alternately enable and disable the TX and RX paths. Even though the transmitter output power

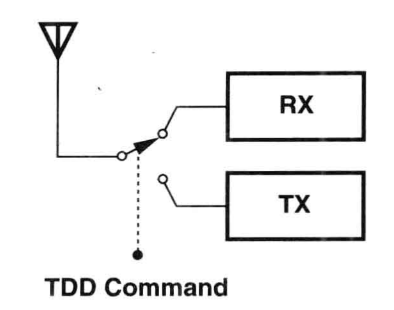


Figure 3.49 *Time-division duplexing.*

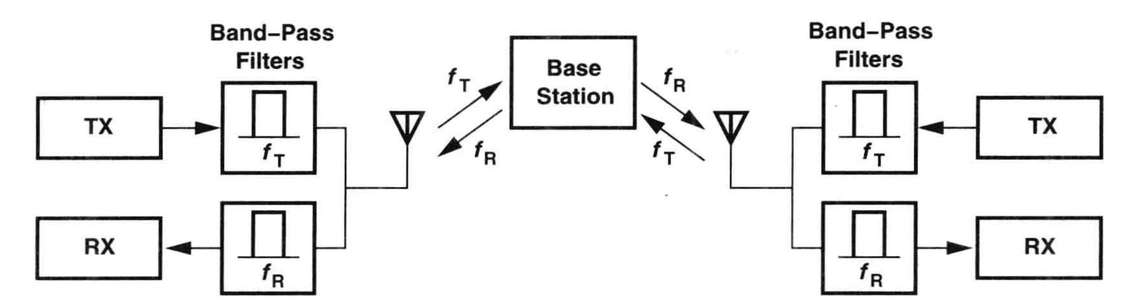


Figure 3.50 *Frequency-division duplexing.*

may be 100 dB above the receiver input signal, the two paths do not interfere because the transmitter is turned off during reception. Furthermore, TDD allows direct ("peer-to-peer") communication between two transceivers, an especially useful feature in short-range, local area network applications. The primary drawback of TDD is that the strong signals generated by all of the nearby mobile transmitters fall in the receive band, thus desensitizing the receiver.

In FDD systems, the two front-end band-pass filters are combined to form a "duplexer filter." While making the receivers immune to the strong signals transmitted by other mobile units, FDD suffers from a number of issues. First, components of the transmitted signal that leak into the receive band are attenuated by typically only about 50 dB (Chapter 4). Second, owing to the trade-off between the loss and the quality factor of filters, the loss of the duplexer is typically quite higher than that of a TDD switch. Note that a loss of, say, 3 dB in the RX path of the duplexer raises the overall noise figure by 3 dB (Chapter 2), and the same loss in the TX path of the filter means that only 50% of the signal power reaches the antenna.

Another issue in FDD is the spectral leakage to adjacent channels in the transmitter output. This occurs when the power amplifier is turned on and off to save energy or when the local oscillator driving the modulator undergoes a transient. By contrast, in TDD such transients can be timed to end before the antenna is switched to the power amplifier output.

Despite the above drawbacks, FDD is employed in many RF systems, particularly in cellular communications, because it isolates the receivers from the signals produced by mobile transmitters.

3.6.2 Frequency-Division Multiple Access

In order to allow simultaneous communication among multiple transceivers, the available frequency band can be partitioned into many channels, each of which is assigned to one user. Called "frequency-division multiple access," (Fig. 3.51), this technique should be familiar within the context of radio and television broadcasting, where the channel assignment does not change with time. In multiple-user, two-way communications, on the other hand, the channel assignment may remain fixed only until the end of the call; after the user hangs up the phone, the channel becomes available to others. Note that in FDMA with FDD, two channels are assigned to each user, one for transmit and another for receive.

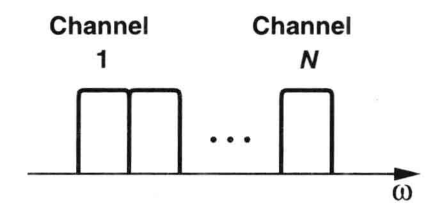


Figure 3.51 *Frequency-division multiple access.*

The relative simplicity of FDMA made it the principal access method in early cellular networks, called "analog FM" systems. However, in FDMA the minimum number of simultaneous users is given by the ratio of the total available frequency band (e.g., 25 MHz in GSM) and the width of each channel (e.g., 200 kHz in GSM), translating to insufficient capacity in crowded areas.

3.6.3 Time-Division Multiple Access

In another implementation of multiple-access networks, the same band is available to each user but at different times ("time-division multiple access"). Illustrated in Fig. 3.52, TDMA periodically enables each of the transceivers for a "time slot" (T_{sl}). The overall period consisting of all of the time slots is called a "frame" (T_F). In other words, every T_F seconds, each user finds access to the channel for T_{sl} seconds.

What happens to the data (e.g., voice) produced by all other users while only one user is allowed to transmit? To avoid loss of information, the data is stored ("buffered") for $T_F - T_{sl}$ seconds and transmitted as a burst during one time slot (hence the term "TDMA burst"). Since buffering requires the data to be in digital form, TDMA transmitters perform A/D conversion on the analog input signal. Digitization also allows speech compression and coding.

TDMA systems have a number of advantages over their FDMA counterparts. First, as each transmitter is enabled for only one time slot in every frame, the power amplifier can be turned off during the rest of the frame, thus saving considerable power. In practice, settling issues require that the PA be turned on slightly before the actual time slot begins. Second, since digitized speech can be compressed in time by a large factor, the required communication bandwidth can be smaller and hence the overall capacity larger. Equivalently, as compressed speech can be transmitted over a shorter time slot, a higher number

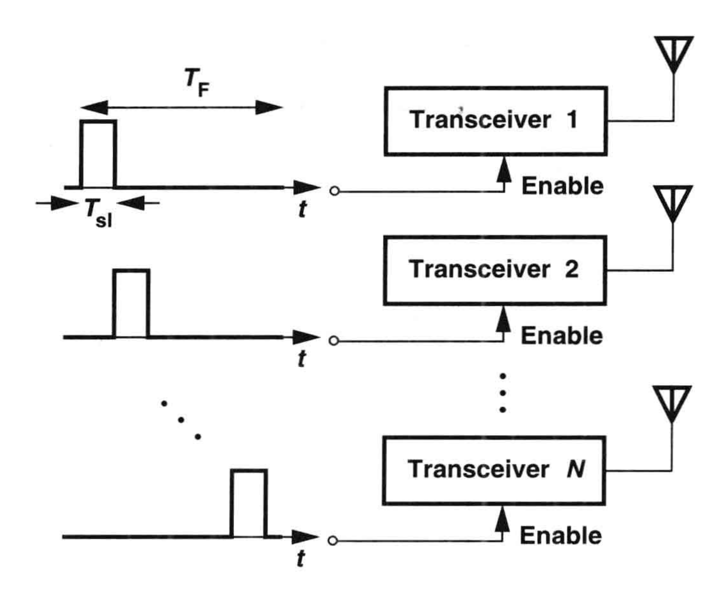


Figure 3.52 *Time-division multiple access.*

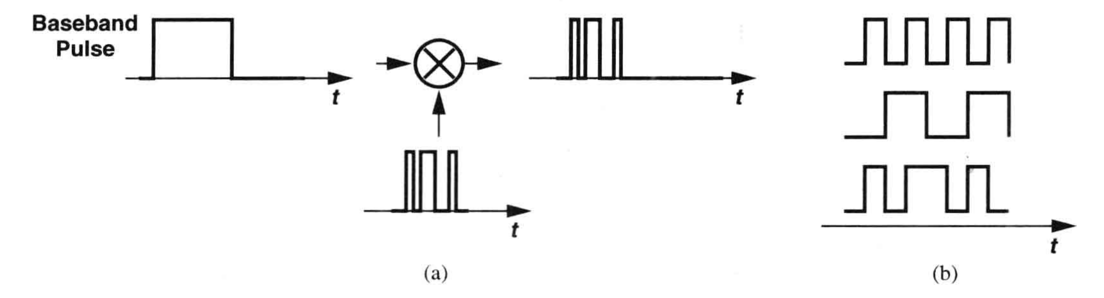
of users can be accommodated in each frame. Third, even with FDD, the TDMA bursts can be timed such that the receive and transmit paths in each transceiver are never enabled simultaneously.

The need for A/D conversion, digital modulation, time slot and frame synchronization, etc., makes TDMA more complex than FDMA. With the advent of VLSI DSPs, however, this drawback is no longer a determining factor. In most actual TDMA systems, a combination of TDMA and FDMA is utilized. In other words, each of the channels depicted in Fig. 3.51 is time-shared among many users.

3.6.4 Code-Division Multiple Access

Our discussion of FDMA and TDMA implies that the transmitted signals in these systems avoid interfering with each other in either the frequency domain or the time domain. In essence, the signals are orthogonal in one of these domains. A third method of multiple access allows complete overlap of signals in both frequency and time, but employs "orthogonal messages" to avoid interference. This can be understood with the aid of an analogy [8]. Suppose many conversations are going on in a crowded party. To minimize crosstalk, different groups of people can be required to speak in different pitches(!) (FDMA), or only one group can be allowed to converse at a time (TDMA). Alternatively, each group can be asked to speak in a *different language*. If the languages are orthogonal (at least in nearby groups) and the voice levels are roughly the same, then each listener can "tune in" to the proper language and receive information while all groups talk simultaneously.

Direct-Sequence CDMA In "code-division multiple access," different languages are created by means of orthogonal digital codes. At the beginning of communication, a certain code is assigned to each transmitter-receiver pair, and each bit of the baseband data is "translated" to that code before modulation. Shown in Fig. 3.53(a) is an example where each baseband pulse is replaced with an 8-bit code by multiplication. A method of



(a) (b)

Figure 3.53 *(a) Encoding operation in DS-CDMA, (b) examples of Walsh code.*

generating orthogonal codes is based on Walsh's recursive equation:

$$W_1 = 0 \tag{3.53}$$

$$W_{2n} = \begin{bmatrix} W_n & W_n \\ W_n & \overline{W_n} \end{bmatrix} \tag{3.54}$$

where $\overline{W_n}$ is derived from W_n by replacing all the entries with their complements. For example,

$$W_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \tag{3.55}$$

Fig. 3.53(b) shows examples of 8-bit Walsh codes (i.e., each row of W_8).

In the receiver, the demodulated signal is decoded by multiplying it by the same Walsh code. In other words, the receiver *correlates* the signal with the code to recover the baseband data.

How is the received data affected when another CDMA signal is present? Suppose two CDMA signals $x_1(t)$ and $x_2(t)$ are received in the same frequency band. Writing the signals as $x_{BB1}(t) \cdot W_1(t)$ and $x_{BB2}(t) \cdot W_2(t)$, where $W_1(t)$ and $W_2(t)$ are Walsh functions, we express the output of the demodulator as $y(t) = [x_{BB1}(t).W_1(t) + x_{BB2}(t).W_2(t)].W_1(t)$. Thus, if $W_1(t)$ and $W_2(t)$ are exactly orthogonal, then $y(t) = x_{BB1}(t) \cdot W_1(t)$. In reality, however, $x_1(t)$ and $x_2(t)$ may experience different delays, leading to corruption of $y(t)$ by $x_{BB2}(t)$. Nevertheless, for long codes the result appears as random noise.

The encoding operation of Fig. 3.53(a) *increases* the bandwidth of the data spectrum by the number of pulses in the code. This may appear in contradiction to our emphasis thus far on spectral efficiency. However, since CDMA allows the widened spectra of many users to fall in the same frequency band (Fig. 3.54), this access technique has no less capacity than do FDMA and TDMA. In fact, CDMA can potentially achieve a higher capacity than the other two [9].

CDMA is a special case of "spread spectrum" (SS) communications, whereby the baseband data of each user is spread over the entire available bandwidth. In this context, CDMA is also called "direct sequence" SS (DS-SS) communication, and the code is called the "spreading sequence" or "pseudo-random noise." To avoid confusion with the baseband data, each pulse in the spreading sequence is called a "chip" and the rate of the sequence

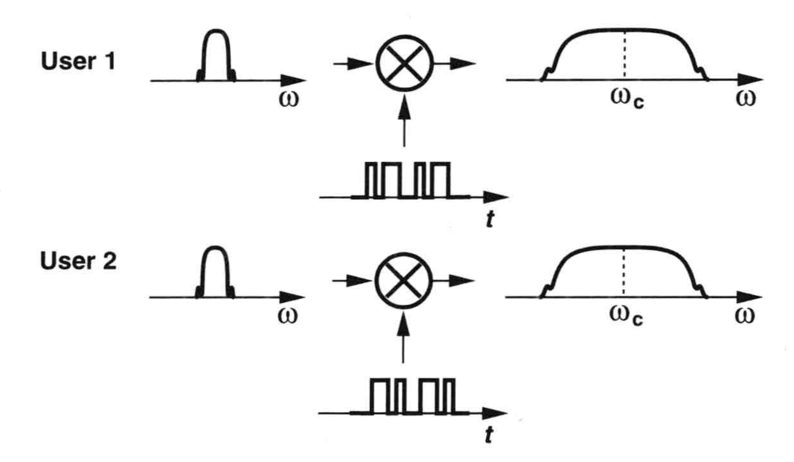


Figure 3.54 *Overlapping spectra in CDMA.*

is called the "chip rate." Thus, the spectrum is spread by the ratio of the chip rate to the baseband bit rate.

It is instructive to reexamine the above RX decoding operation from a spread spectrum point of view (Fig. 3.55). Upon multiplication by $W_1(t)$, the desired signal is "despread," with its bandwidth returning to the original value. The unwanted signal, on the other hand, remains spread even after multiplication because of its low correlation with $W_1(t)$. For a large number of users, the spread spectra of unwanted signals can be viewed as white Gaussian noise.

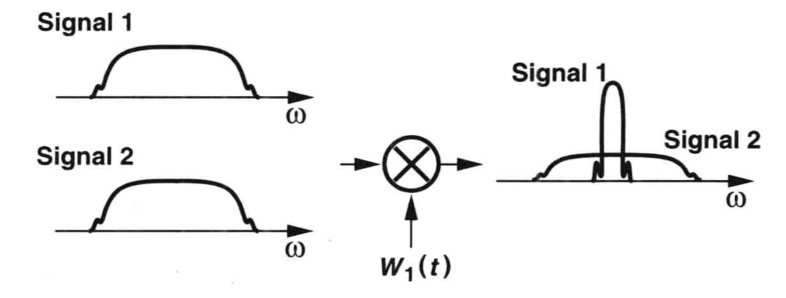


Figure 3.55 *Despreading operation in a CDMA receiver.*

An important feature of CDMA is its soft capacity limit [7]. While in FDMA and TDMA the maximum number of users is fixed once the channel width or the time slots are defined, in CDMA increasing the number of users only gradually (linearly) raises the noise floor [7].

A critical issue in DS-CDMA is power control. Suppose, as illustrated in Fig. 3.56, the desired signal power received at a point is much lower than that of an unwanted transmitter,[7] for example because the latter is at a shorter distance. Even after despreading, the strong interferer greatly raises the noise floor, degrading the reception of the desired signal. For multiple users, this means that one high-power transmitter can virtually halt

7. This situation arises in our party analogy if two people speak much more loudly than others. Even with different languages, communication becomes difficult.

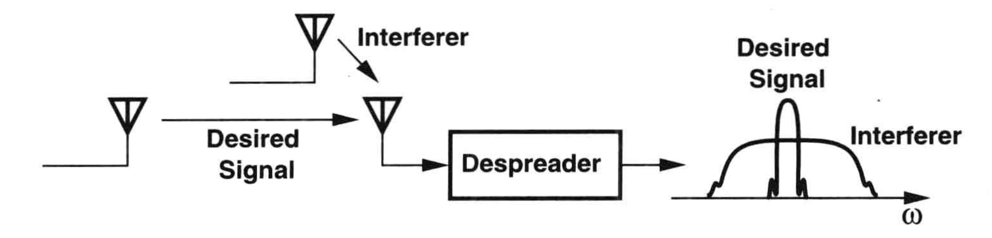


Figure 3.56 *Near/far effect in CDMA.*

communications among others, a problem much less serious in FDMA and TDMA. This is called the "near/far effect." For this reason, when many CDMA transmitters communicate with a receiver, they must adjust their output power such that the receiver senses roughly equal signal levels. To this end, the receiver monitors the signal strength corresponding to each transmitter and periodically sends a power adjustment request to each one. Since in a cellular system users communicate through the base station, rather than directly, the latter must handle the task of power control. The received signal levels are controlled to be typically within 1 dB of each other.

While adding complexity to the system, power control generally reduces the average power dissipation of the mobile unit. To understand this, note that without such control, the mobile must *always* transmit enough power to be able to communicate with the base station, whether path loss and fading are significant or not. Thus, even when the channel has minimum attenuation, the mobile unit produces the maximum output power. With power control, on the other hand, the mobile can transmit at low levels whenever the channel conditions improve. This also reduces the average interference seen by other users.

Unfortunately, power control also dictates that the receive and transmit paths of the mobile phone operate concurrently.[8] As a consequence, CDMA mobile phones must deal with the leakage of the TX signal to the RX (Chapter 4).

Frequency-Hopping CDMA Another type of CDMA that has begun to appear in RF communications is "frequency hopping" (FH). Illustrated in Fig. 3.57, this access technique can be viewed as FDMA with pseudo-random channel allocation. The carrier frequency in

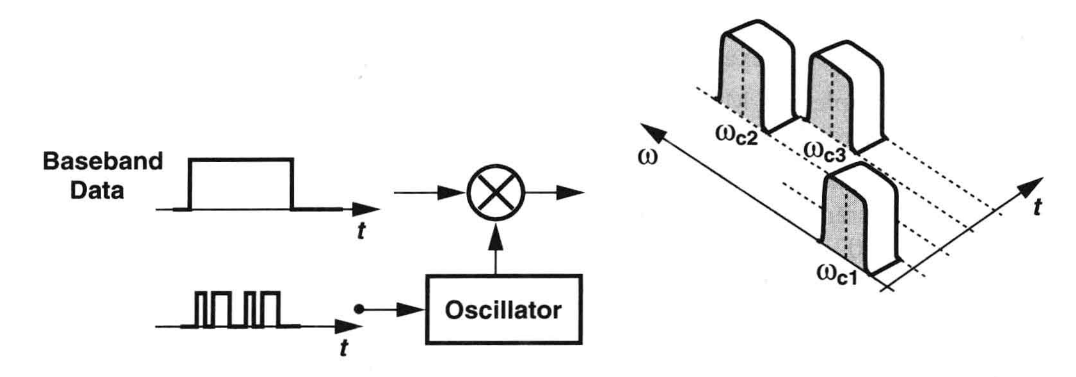


Figure 3.57 *Frequency-hopping CDMA.*

8. If a vehicle moves at a high speed or in an area with tall buildings, the power received by the base station from it can vary rapidly, requiring continuous feedback.

each transmitter is "hopped" according to a chosen code (similar to the spreading codes in DS-CDMA). Thus, even though the short-term spectrum of a transmitter may overlap with those of others, the overall trajectory of the spectrum, i.e., the PN code, distinguishes each transmitter from others. Nevertheless, occasional overlap of the spectra raises the probability of error.

Due to rare overlap of spectra, frequency hopping is somewhat similar to FDMA and hence more tolerant of different received power levels than is direct-sequence CMDA. However, FH may require relatively fast settling in the control loop of the oscillator shown in Fig. 3.57, an important design issue studied in Chapter 10.

3.7 WIRELESS STANDARDS

Our study of wireless communication systems thus far indicates that making a phone call or sending data entails a great many complex operations in both analog and digital domains. Furthermore, nonidealities such as noise and interference require precise specification of each system parameter, e.g., SNR, BER, occupied bandwidth, and tolerance of interferers. A "wireless standard" defines the essential functions and specifications that govern the design of the transceiver, including its baseband processing. Anticipating various operating conditions, each standard fills a relatively large document while still leaving some of the dependent specifications for the designer to choose. For example, a standard may specify the sensitivity but not the noise figure.

Before studying various wireless standards, we briefly consider some of the common specifications that standards quantify:

1. Frequency Bands and Channelization. Each standard performs communication in an allocated frequency band. For example, Bluetooth uses the industrial-scientific-medical (ISM) band from 2.400 GHz to 2.480 GHz. The band consists of "channels," each of which carries the information for one user. For example, Bluetooth incorporates a channel of 1 MHz, allowing at most 80 users.
2. Data Rate(s). The standard specifies the data rate(s) that must be supported. Some standards support a constant data rate, whereas others allow a variable data rate so that, in the presence of high signal attenuation, the communication is sustained but at a low speed. For example, Bluetooth specifies a data rate of 1 Mb/s.
3. Antenna Duplexing Method. Most cellular phone systems incorporate FDD, and other standards employ TDD.
4. Type of Modulation. Each standard specifies the modulation scheme. In some cases, different modulation schemes are used for different data rates. For example, IEEE802.11a/g utilizes 64QAM for its highest rate (54 Mb/s) in the presence of good signal conditions, but binary PSK for the lowest rate (6 Mb/s).
5. TX Output Power. The standard specifies the power level(s) that the TX must produce. For example, Bluetooth transmits 0 dBm. Some standards require a variable output level to save battery power when the TX and RX are close to each other and/or to avoid near/far effects.
6. TX EVM and Spectral Mask. The signal transmitted by the TX must satisfy several requirements in addition to the power level. First, to ensure acceptable signal

quality, the EVM is specified. Second, to guarantee that the TX out-of-channel emissions remain sufficiently small, a TX "spectral mask" is defined. As explained in Section 3.4, excessive PA nonlinearity may violate this mask. Also, the standard poses a limit on other unwanted transmitted components, e.g., spurs and harmonics.

7. RX Sensitivity. The standard specifies the acceptable receiver sensitivity, usually in terms of a maximum bit error rate, BER_{max}. In some cases, the sensitivity is commensurate with the data rate, i.e., a higher sensitivity is stipulated for lower data rates.

8. RX Input Level Range. The desired signal sensed by a receiver may range from the sensitivity level to a much larger value if the RX is close to the TX. Thus, the standard specifies the desired signal range that the receiver must handle with acceptable noise or distortion.

9. RX Tolerance to Blockers. The standard specifies the largest interferer that the RX must tolerate while receiving a small desired signal. This performance is typically defined as illustrated in Fig. 3.58. In the first step, a modulated signal is applied at the "reference" sensitivity level and the BER is measured to remain below BER_{max} [Fig. 3.58(a)]. In the second step, the signal level is raised by 3 dB and a blocker is added to the input and its level is gradually raised. When the blocker reaches the specified level, the BER must not exceed BER_{max} [Fig. 3.58(b)]. This test reveals the compression behavior and phase noise of the receiver. The latter is described in Chapter 8.

   Many standards also stipulate an intermodulation test. For example, as shown in Fig. 3.59, two blockers (one modulated and another not) are applied along with the desired signal at 3 dB above the sensitivity level. The receiver BER must not exceed BER_{max} as the level of the two blockers reaches the specified level.

In this section, we study a number of wireless standards. In the case of cellular standards, we focus on the "mobile station" (the handset).

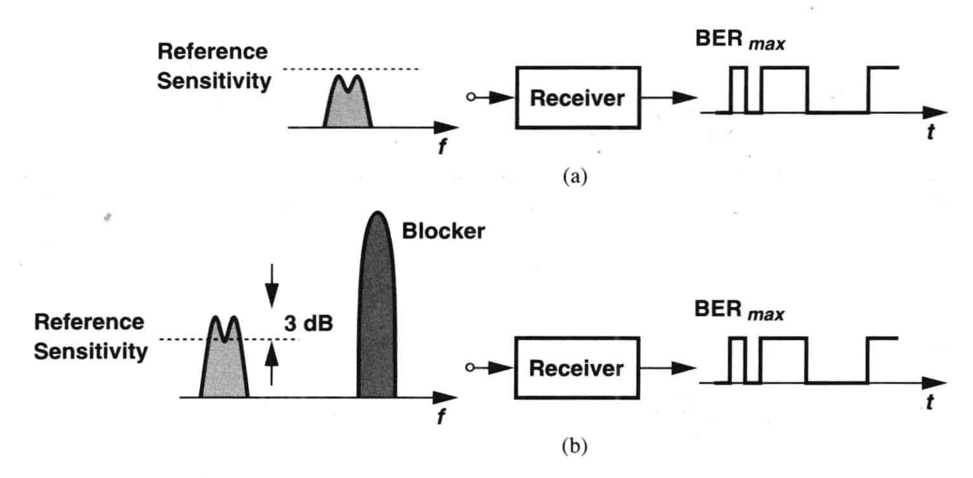


Figure 3.58 *Test of a receiver with (a) desired signal at reference sensitivity and (b) desired signal 3 dB above reference sensitivity along with a blocker.*

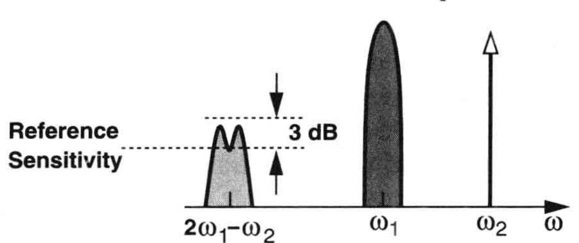


Figure 3.59 *Intermodulation test.*

3.7.1 GSM

The Global System for Mobile Communication (GSM) was originally developed as a unified wireless standard for Europe and became the most widely-used cellular standard in the world. In addition to voice, GSM also supports the transmission of data.

The GSM standard is a TDMA/FDD system with GMSK modulation, operating in different bands and accordingly called GSM900, GSM1800 (also known as DCS1800), and GSM1900 (also known as PCS1900). Figure 3.60 shows the TX and RX bands. Accommodating eight time-multiplexed users, each channel is 200 kHz wide, and the data rate per user is 271 kb/s. The TX and RX time slots are offset (by about 1.73 ms) so that the two paths do not operate simultaneously. The total capacity of the system is given by the number of channels in the 25-MHz bandwidth and the number of users is per channel, amounting to approximately 1,000.

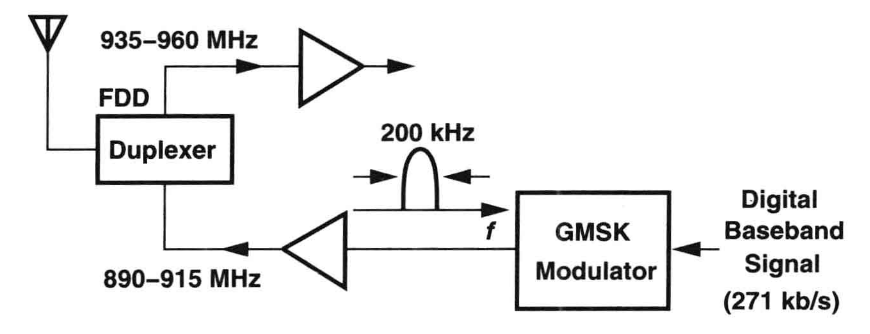


Figure 3.60 *GSM air interface.*

Example 3.10

GSM specifies a receiver sensitivity of -102 dBm.[9] The detection of GMSK with acceptable bit error rate (10^{-3}) requires an SNR of about 9 dB. What is the maximum allowable RX noise figure?

Solution:

We have from Chapter 2

$$NF = 174\text{ dBm/Hz} - 102\text{ dBm} - 10\log(200\text{ kHz}) - 9\text{ dB} \tag{3.56}$$
$$\approx 10\text{ dB}. \tag{3.57}$$

9. The sensitivity in GSM1800 is -101 dBm.

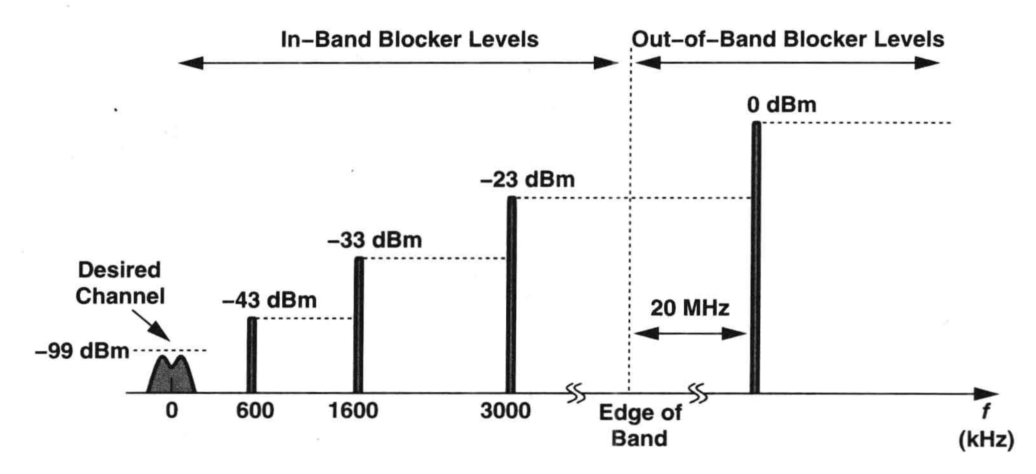


Figure 3.61 *GSM receiver blocking test. (The desired channel center frequency is denoted by 0 for simplicity.)*

Blocking Requirements GSM also specifies blocking requirements for the receiver. Illustrated in Fig. 3.61, the blocking test applies the desired signal at 3 dB above the sensitivity level along with a single (unmodulated) *tone* at discrete increments of 200 kHz from the desired channel. (Only one blocker is applied at a time.)[10] The tolerable in-band blocker level jumps to −33 dBm at 1.6 MHz from the desired channel and to −23 dBm at 3 MHz. The out-of-band blocker can reach 0 dBm beyond a 20-MHz guard band from the edge of the RX band. With the blocker levels shown in Fig. 3.61, the receiver must still provide the necessary BER.

Example 3.11

How must the receiver P_{1dB} be chosen to satisfy the above blocking tests?

Solution:

Suppose the receiver incorporates a front-end filter and hence provides sufficient attenuation if the blocker is applied outside the GSM band. Thus, the largest blocker level is equal to −23 dBm (at or beyond 3-MHz offset), demanding a P_{1dB} of roughly −15 dBm to avoid compression. If the front-end filter does not attenuate the out-of-band blocker adequately, then a higher P_{1dB} is necessary.

If the receiver P_{1dB} is determined by the blocker levels beyond 3-MHz offset, why does GSM specify the levels at smaller offsets? Another receiver imperfection, namely, the phase noise of the oscillator, manifests itself here and is discussed in Chapter 8.

Since the blocking requirements of Fig. 3.61 prove difficult to fulfill in practice, GSM stipulates a set of "spurious response exceptions," allowing the blocker level at six in-band

10. This mask and others described in this section symmetrically extend to the left.

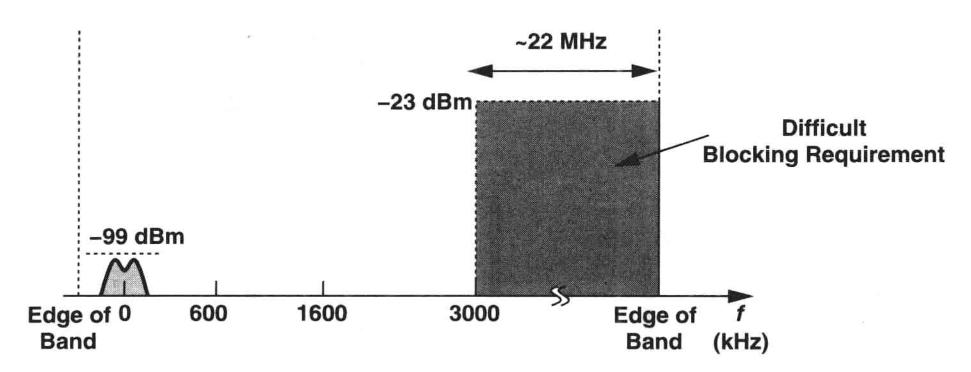


Figure 3.62 *Worst-case channel for GSM blocking test.*

frequencies and 24 out-of-band frequencies to be relaxed to −43 dBm.[11] Unfortunately, these exceptions do not ease the compression and phase noise requirements. For example, if the desired channel is near one edge of the band (Fig. 3.62), then about 100 channels reside above 3-MHz offset. Even if six of these channels are excepted, each of the remaining can contain a −23 dBm blocker.

Intermodulation Requirements Figure 3.63 depicts the IM test specified by GSM. With the desired channel 3 dB above the reference sensitivity level, a tone and a modulated signal are applied at 800-kHz and 1.6-MHz offset, respectively. The receiver must satisfy the required BER if the level of the two interferers is as high as −49 dBm.

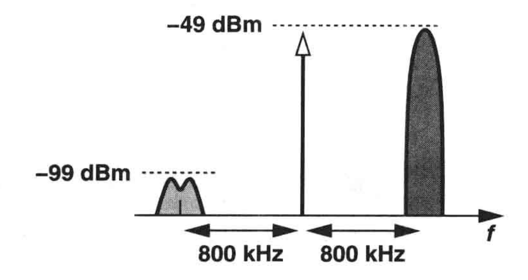


Figure 3.63 *GSM intermodulation test.*

Example 3.12

Estimate the receiver IP_3 necessary for the above test.

Solution:

For an acceptable BER, an SNR of 9 dB is required, i.e., the total noise in the desired channel must remain below −108 dBm. In this test, the signal is corrupted by both the receiver noise and the intermodulation. If, from Example 3.10, we assume NF = 10 dB, then the total RX noise in 200 kHz amounts to −111 dBm. Since the maximum tolerable

11. In GSM1800 and GSM1900, 12 in-band exceptions are allowed.

Example 3.12 *(Continued)*

noise is -108 dBm, the intermodulation can contribute at most 3 dB of corruption. In other words, the IM product of the two interferers must have a level of -111 dBm so that, along with an RX noise of -111 dBm, it yields a total corruption of -108 dBm. It follows from Chapter 2 that

$$\mathrm{IIP}_3 = \frac{-49\ \mathrm{dBm} - (-111\ \mathrm{dBm})}{2} + (-49\ \mathrm{dBm}) \tag{3.58}$$

$$= -18\ \mathrm{dBm}. \tag{3.59}$$

In Problem 3.2, we recompute the IIP_3 if the noise figure is lower than 10 dB.

We observe from this example and Example 3.11 that the receiver linearity in GSM is primarily determined by the single-tone blocking requirements rather than the intermodulation specification.

Adjacent-Channel Interference A GSM receiver must withstand an adjacent-channel interferer 9 dB above the desired signal or an alternate-adjacent channel interferer 41 dB above the desired signal (Fig. 3.64). In this test, the desired signal is 20 dB higher than the sensitivity level. As explained in Chapter 4, the relatively relaxed adjacent-channel requirement facilitates the use of certain receiver architectures.

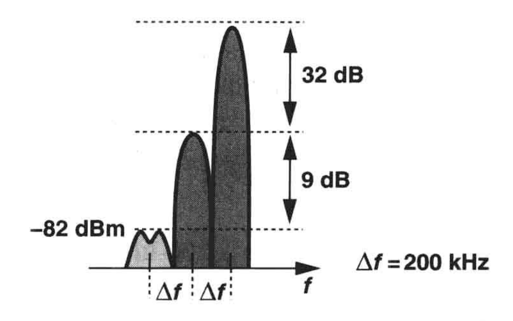


Figure 3.64 *GSM adjacent-channel test.*

TX Specifications A GSM (mobile) transmitter must deliver an output power of at least 2 W (+33 dBm) in the 900-MHz band or 1 W in the 1.8-GHz band. Moreover, the output power must be adjustable in steps of 2 dB from +5 dBm to the maximum level, allowing adaptive power control as the mobile comes closer to or goes farther from the base station.

The output spectrum produced by a GSM transmitter must satisfy the "mask" shown in Fig. 3.65, dictating that GMSK modulation be realized with an accurate modulation index and well-controlled pulse shaping. Also, the rms phase error of the output signal must remain below 5°.

A stringent specification in GSM relates to the maximum noise that the TX can emit in the receive band. As shown in Fig. 3.66, this noise level must be less than -129 dBm/Hz

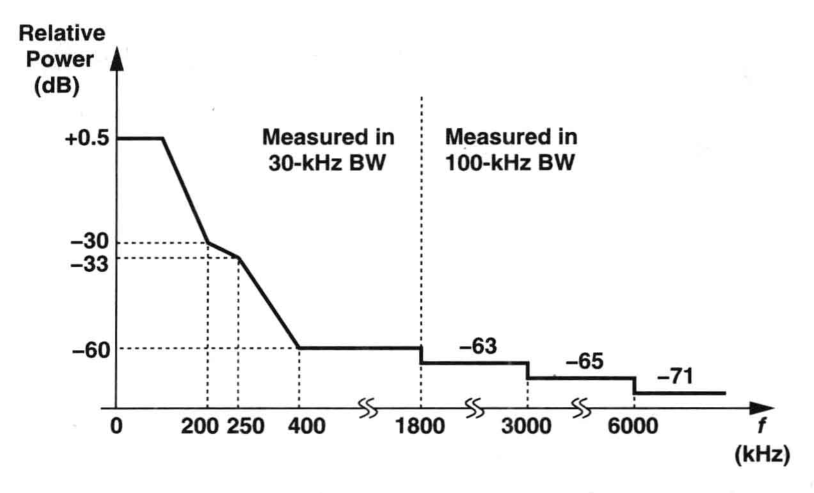


Figure 3.65 *GSM transmission mask for the 900-MHz band.*

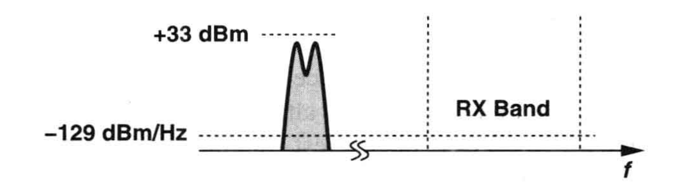


Figure 3.66 *GSM transmitter noise in receive band.*

so that a transmitting mobile station negligibly interferes with a receiving mobile station in its close proximity. The severity of this requirement becomes obvious if the noise bound is normalized to the TX output power of +33 dBm, yielding a relative noise floor of −162 dBc/Hz. As explained in Chapter 4, this specification makes the design of GSM transmitters quite difficult.

EDGE To accommodate higher data rates in a 200-kHz channel, the GSM standard has been extended to "Enhanced Data Rates for GSM Evolution" (EDGE). Achieving a rate of 384 kb/s, EDGE employs "8-PSK" modulation, i.e., phase modulation with eight phase values given by $k\pi/4$, k = 0-7. Figure 3.67 shows the signal constellation. EDGE is considered a "2.5th-generation" (2.5G) cellular system.

The use of 8-PSK modulation entails two issues. First, to confine the spectrum to 200 kHz, a "linear" modulation with baseband pulse shaping is necessary. In fact, the constellation of Fig. 3.67 can be viewed as that of two QPSK waveforms, one rotated by 45° with respect to the other. Thus, two QPSK signals with pulse shaping (Chapter 4) can be generated and combined to yield the 8-PSK waveform. Pulse shaping, however, leads to a variable envelope, necessitating a linear power amplifier. In other words, a GSM/EDGE transmitter can operate with a nonlinear (and hence efficient) PA in the GSM mode but must switch to a linear (and hence inefficient) PA in the EDGE mode.

The second issue concerns the detection of the 8-PSK signal in the receiver. The closely-spaced points in the constellation require a higher SNR than, say, QPSK does. For BER = 10^{-3}, the former dictates an SNR of 14 dB and the latter, 7 dB.

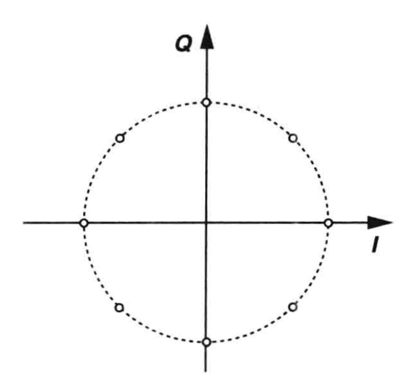


Figure 3.67 *Constellation of 8-PSK (used in EDGE).*

3.7.2 IS-95 CDMA

A wireless standard based on direct-sequence CMDA has been proposed by Qualcomm, Inc., and adopted for North America as IS-95. Using FDD, the air interface employs the transmit and receive bands shown in Fig. 3.68. In the mobile unit, the 9.6 kb/s baseband data is spread to 1.23 MHz and subsequently modulated using OQPSK. The link from the base station to the mobile unit, on the other hand, incorporates QPSK modulation. The logic is that the mobile must use a power-efficient modulation scheme (Chapter 3), whereas the base station transmits many channels simultaneously and must therefore employ a linear power amplifier regardless of the type of modulation. In both directions, IS-95 requires *coherent* detection, a task accomplished by transmitting a relatively strong "pilot tone" (e.g., unmodulated carrier) at the beginning of communication to establish phase synchronization.

In contrast to the other standards studied above, IS-95 is substantially more complex, incorporating many techniques to increase the capacity while maintaining a reasonable signal quality. We briefly describe some of the features here. For more details, the reader is referred to [8, 10, 11].

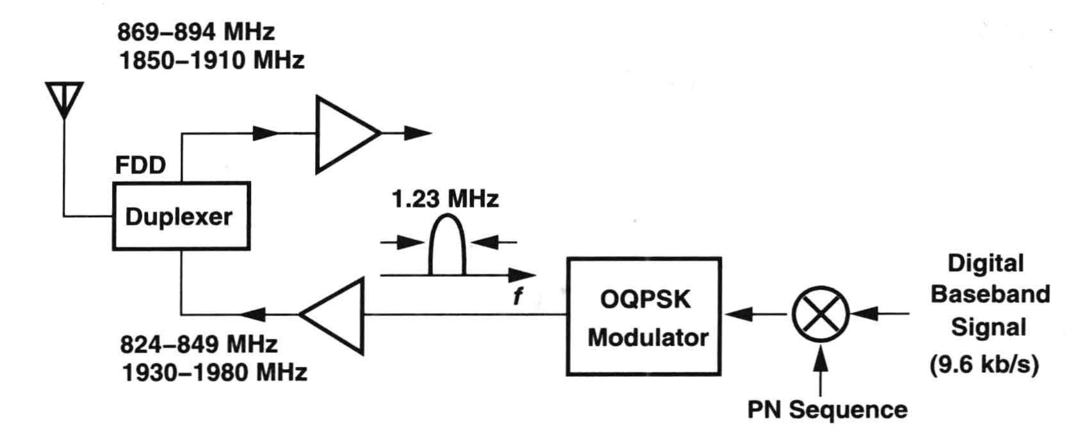


Figure 3.68 *IS-95 air interface.*

Power Control As mentioned in Section 3.6.4, in CDMA the power levels received by the base station from various mobile units must differ by no more than approximately 1 dB. In IS-95, the output power of each mobile is controlled by an open-loop procedure at the beginning of communication so as to perform a rough but fast adjustment. Subsequently, the power is set more accurately by a closed-loop method. For open-loop control, the mobile measures the signal power it *receives* from the base station and adjusts its transmitted power so that the sum of the two (in dB) is approximately -73 dBm. If the receive and transmit paths entail roughly equal attenuation, k dB, and the power transmitted by the base station is P_{bs}, then the mobile output power P_m satisfies the following: $P_{bs} - k + P_m = -73$ dBm. Since the power received by the base station is $P_m - k$, we have $P_m - k = -73$ dBm $-P_{bs}$, a well-defined value because P_{bs} is usually fixed. The mobile output power can be varied by approximately 85 dB in a few microseconds.

Closed-loop power control is also necessary because the above assumption of equal loss in the transmit and receive paths is merely an approximation. In reality, the two paths may experience different fading because they operate in different frequency bands. To this end, the base station measures the power level received from the mobile unit and sends a feedback signal requesting power adjustment. This command is transmitted once every 1.25 ms to ensure timely adjustment in the presence of rapid fading.

Frequency and Time Diversity Recall from Section 3.5 that multipath fading is often frequency-selective, causing a notch in the channel transfer function that can be several kilohertz wide. Since IS-95 spreads the spectrum to 1.23 MHz, it provides frequency diversity, exhibiting only 25% loss of the band for typical delay spreads [8].

IS-95 also employs time diversity to use multipath signals to advantage. This is accomplished by performing correlation on *delayed replicas* of the received signal (Fig. 3.69). Called a "rake receiver," such a system combines the delayed replicas with proper weighting factors, α_j, to obtain the maximum signal-to-noise ratio at the output. That is, if the output of one correlator is corrupted, then the corresponding weighting factor is reduced and vice versa.

Rake receivers are a unique feature of CDMA. Since the chip rate is much higher than the fading bandwidth, and since the spreading codes are designed to have negligible correlation for delays greater than a chip period, multipath effects do not introduce intersymbol interference. Thus, each correlator can be synchronized to one of the multipath signals.

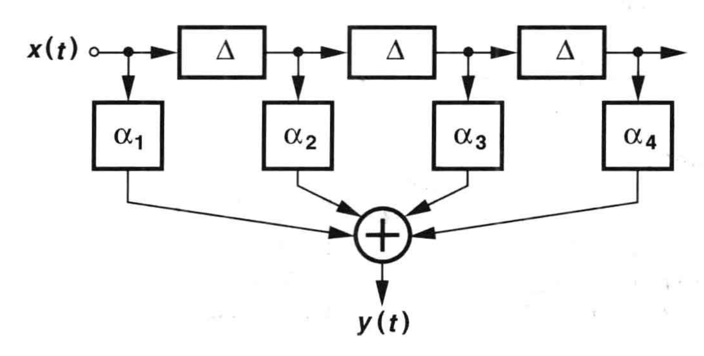


Figure 3.69 *Rake receiver.*

Variable Coding Rate The variable rate of information in human speech can be exploited to lower the average number of transmitted bits per second. In IS-95, the data rate can vary in four discrete steps: 9600, 4800, 2400, and 1200 b/s. This arrangement allows buffering slower data such that the transmission still occurs at 9600 b/s but for a proportionally shorter duration. This approach further reduces the average power transmitted by the mobile unit, both saving battery and lowering interference seen by other users.

Soft Hand-off Recall from Section 3.5 that when the mobile unit is assigned a different base station, the call may be dropped if the channel center frequency must change (e.g., in IS-54 and GSM). In CDMA, on the other hand, all of the users in one cell communicate on the same channel. Thus, as the mobile unit moves farther from one base station and closer to another, the signal strength corresponding to *both* stations can be monitored by means of a rake receiver. When it is ascertained that the nearer base station has a sufficiently strong signal, the hand-off is performed. Called "soft hand-off," this method can be viewed as a "make-before-break" operation. The result is lower probability of dropping calls during hand-off.

3.7.3 Wideband CDMA

As a third-generation cellular system, wideband CDMA extends the concepts realized in IS-95 to achieve a higher data rate. Using BPSK (for uplink) and QPSK (for downlink) in a nominal channel bandwidth of 5 MHz, WCDMA achieves a rate of 384 kb/s.

Several variants of WCDMA have been deployed in different graphical regions. In this section, we study "IMT-2000" as an example. Figure 3.70 shows the air interface of IMT-2000, indicating a total bandwidth of 60 MHz. Each channel can accommodate a data rate of 384 kb/s in a (spread) bandwidth of 3.84 MHz; but, with "guard bands" included, the channel spacing is 5 MHz. The mobile station employs BPSK modulation for data and QPSK modulation for spreading.

Transmitter Requirements The TX must deliver an output power ranging from -49 dBm to $+24$ dBm.[12] The wide output dynamic range makes the design of WCDMA transmitters and, specifically, power amplifiers, difficult. In addition, the TX incorporates baseband pulse shaping so as to tighten the output spectrum, calling for a linear PA.

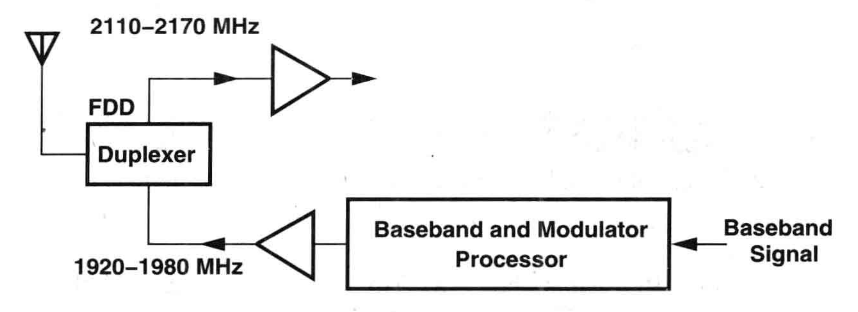


Figure 3.70 *IMT-2000 air interface.*

12. The PA may need to deliver about +27 dBm to account for the loss of the duplexer.

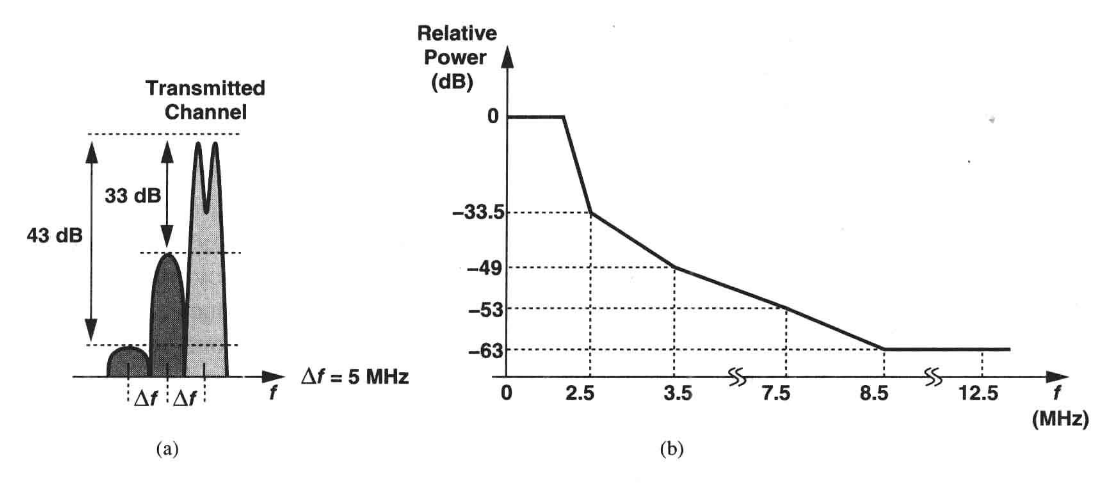


Figure 3.71 *Transmitter (a) adjacent-channel power test, and (b) emission mask in IMT-2000.*

IMT-2000 stipulates two sets of specifications to quantify the limits on the out-of-channel emissions: (1) the adjacent and alternate adjacent channel powers must be 33 dB and 43 dB below the main channel, respectively [Fig. 3.71(a)]; (2) the emissions measured in a 30-kHz bandwidth must satisfy the TX mask shown in Fig. 3.71(b).

The transmitter must also coexist harmoniously with the GSM and DCS1800 standards. That is, the TX power must remain below −79 dBm in a 100-kHz bandwidth in the GSM RX band (935 MHz-960 MHz) and below −71 dBm in a 100-kHz bandwidth in the DCS1800 RX band (1805 MHz-1880 MHz).

Receiver Requirements The receiver reference sensitivity is −107 dBm. As with GSM, IMT-2000 receivers must withstand a sinusoidal blocker whose amplitude becomes larger at greater frequency offsets. Unlike GSM, however, IMT-2000 requires the sinusoidal test for only *out-of-band* blocking [Fig. 3.72(a)]. The P_{1dB} necessary here is more relaxed than that in GSM; e.g., at 85 MHz outside the band, the RX must tolerate a tone of −15 dBm.[13] For in-band blocking, IMT-2000 provides the two tests shown in Fig. 3.72(b). Here, the blocker is modulated such that it behaves as another WCDMA channel, thus causing both compression and cross modulation.

Example 3.13

Estimate the required P_{1dB} of a WCDMA receiver satisfying the in-band test of Fig. 3.72(b).

Solution:

To avoid compression, P_{1dB} must be 4 to 5 dB higher than the blocker level, i.e., $P_{1dB} \approx -40$ dBm. To quantify the corruption due to cross modulation, we return to our derivation in Chapter 2. For a sinusoid $A_1 \cos \omega_1 t$ and an amplitude-modulated blocker

13. However, if the TX leakage is large, the RX linearity must be quite higher.

Example 3.13 (*Continued*)

$A_2(1 + m\cos\omega_m t)\cos\omega_2 t$, cross modulation appears as

$$y(t) = \left[\alpha_1 A_1 + \frac{3}{2}\alpha_3 A_1 A_2^2\left(1 + \frac{m^2}{2} + \frac{m^2}{2}\cos 2\omega_m t + 2m\cos\omega_m t\right)\right]\cos\omega_1 t + \cdots \quad (3.60)$$

For the case at hand, both channels contain modulation and we make the following assumptions: (1) the desired channel and the blocker carry the same amplitude modulation and are respectively expressed as $A_1(1+m\cos\omega_{m1}t)\cos\omega_1 t$ and $A_2(1+m\cos\omega_{m2}t)\cos\omega_2 t$; (2) the envelope varies by a moderate amount and hence $m^2/2 \ll 2m$. The effect of third-order nonlinearity can then be expressed as

$$y(t) = \left[\alpha_1 A_1(1+m\cos\omega_{m1}t)+\frac{3}{2}\alpha_3 A_1(1+m\cos\omega_{m1}t)A_2^2 \times (1+2m\cos\omega_{m2}t)\right]\cos\omega_1 t+\cdots \quad (3.61)$$

$$= \left[\alpha_1 A_1(1+m\cos\omega_{m1}t)+\frac{3}{2}\alpha_3 A_1 A_2^2(1+m\cos\omega_{m1}t+2m\cos\omega_{m2}t +2m^2\cos\omega_{m1}t\cos\omega_{m2}t)\right]\cos\omega_1 t+\cdots \quad (3.62)$$

For the corruption to be negligible, the average power of the second term in the square brackets must remain much less than that of the first:

$$\frac{\left(\frac{3}{2}\alpha_3 A_1 A_2^2\right)^2(1 + m^2 + 4m^2 + 4m^4)}{(\alpha_1 A_1)^2(1 + m^2)} \ll 1. \quad (3.63)$$

Setting this ratio to -15 dB ($= 0.0316$) and neglecting the powers of m, we have

$$\frac{\frac{3}{2}|\alpha_3|A_2^2}{|\alpha_1|} = 0.178. \quad (3.64)$$

Since $A_{1dB} = \sqrt{0.145|\alpha_1/\alpha_3|}$ and hence $|\alpha_3/\alpha_1| = 0.145/A_{1dB}^2$,

$$A_{1dB} = 1.1\,A_2. \quad (3.65)$$

That is, the input compression point must exceed A_2 ($= -44$ dBm) by about 1 dB. Thus, compression is slightly more dominant than cross modulation in this test.

A sensitivity level of -107 dBm with a signal bandwidth as wide as 3.84 MHz may appear very impressive. In fact, since $10\log(3.84\text{ MHz}) \approx 66$ dB, it seems that the sum of the receiver noise figure and the required SNR must not exceed 174 dBm/Hz $-$ 66 dB $-$ 107 dBm $=$ 1 dB! However, recall that CDMA *spreads* a lower bit rate (e.g., 384 kb/s)

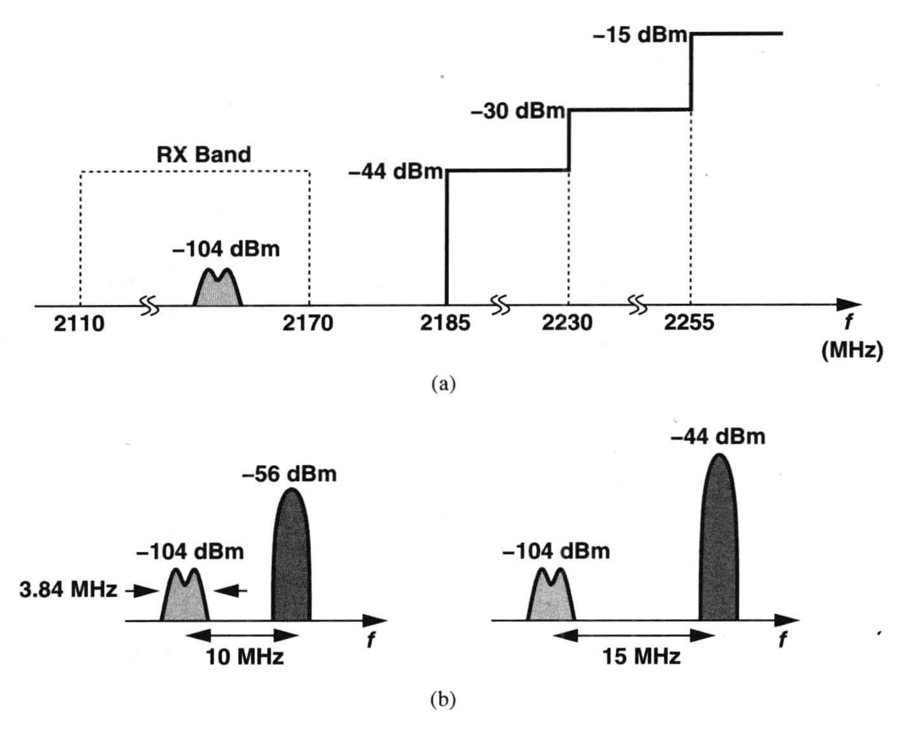


Figure 3.72 *IMT-2000 receiver (a) blocking mask using a tone and (b) blocking test using a modulated interferer.*

by a factor, thus benefitting from the spreading gain after the despreading operation in the receiver. That is, the NF is relaxed by a factor equal to the spreading gain.

IMT-2000 also specifies an intermodulation test. As depicted in Fig. 3.73, a tone and a modulated signal, each at -46 dBm, are applied in the adjacent and alternate adjacent channels, respectively, while the desired signal is at -104 dBm. In Problem 3.3, we repeat Example 3.12 for WCDMA to determine the required IP_3 of the receiver.

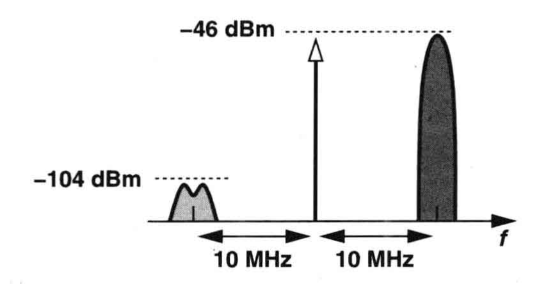


Figure 3.73 *IMT-2000 intermodulation test.*

Figure 3.74 illustrates the adjacent-channel test stipulated by IMT-2000. With a level of -93 dBm for the desired signal, the adjacent channel can be as high as -52 dBm. As explained in Chapter 4, this specification requires a sharp roll-off in the frequency response of the baseband filters.

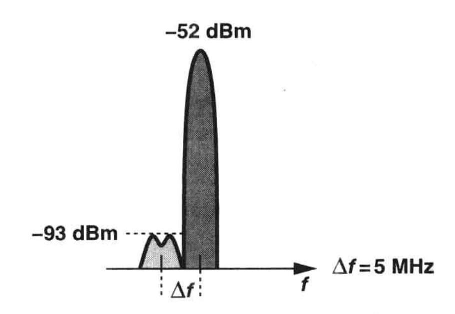


Figure 3.74 *IMT-2000 receiver adjacent-channel test.*

3.7.4 Bluetooth

The Bluetooth standard was originally conceived as a low-cost solution for short-range, moderate-rate data communication. In fact, it was envisioned that the transceiver would perform modulation and demodulation in the analog domain, requiring little digital signal processing. In practice, however, some attributes of the standard have made the analog implementations difficult, calling for substantial processing in the digital domain. Despite these challenges, Bluetooth has found its place in the consumer market, serving in such short-reach applications as wireless headsets, wireless keyboards, etc.

Figure 3.75 shows the Bluetooth air interface, indicating operation in the 2.4-GHz ISM band. Each channel carries 1 Mb/s, occupies 1 MHz, and has a carrier frequency equal to $(2402 + k)$ MHz, for $k = 0, \cdots, 78$. To comply with out-of-band emission requirements of various countries, the first 2 MHz and last 3.5 MHz of the ISM band are saved as "guard bands" and not used.

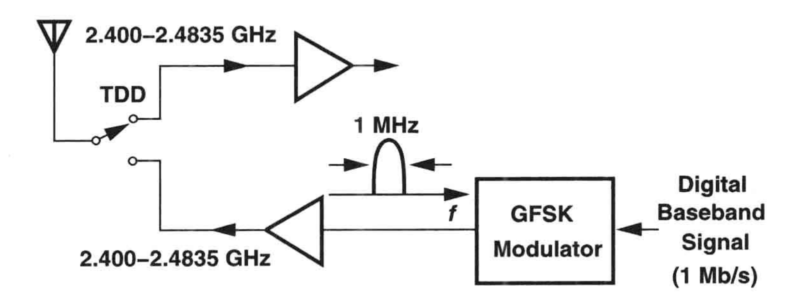


Figure 3.75 *Bluetooth air interface.*

Transmitter Characteristics The 1-Mb/s baseband data is applied to a Gaussian Frequency Shift Keying (GFSK) modulator. GFSK can be viewed as GMSK with a modulation index of 0.28 to 0.35. As explained in Section 3.3.4, GMSK modulation can be realized by a Gaussian filter and a VCO (Fig. 3.76).

The output can be expressed as

$$x_{TX}(t) = A\cos[\omega_c t + m\int x_{BB}(t) * h(t)dt], \tag{3.66}$$

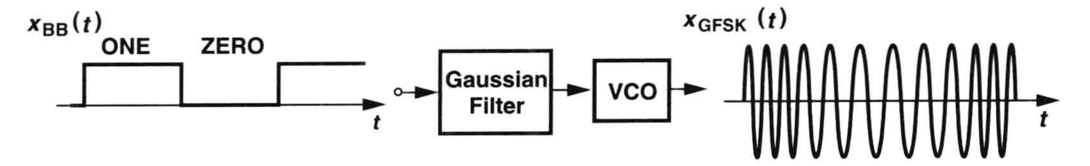


Figure 3.76 *GFSK modulation using a VCO.*

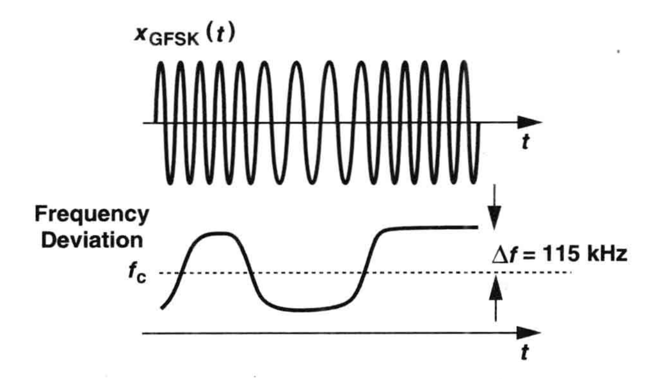


Figure 3.77 *Frequency deviation in Bluetooth.*

where $h(t)$ denotes the impulse response of the Gaussian filter. As shown in Fig. 3.77, Bluetooth specifies a minimum frequency deviation of 115 kHz in the carrier. The peak frequency deviation, Δf, is obtained as the maximum difference between the instantaneous frequency, f_{inst}, and the carrier frequency. Differentiating the argument of the cosine in (3.66) yields f_{inst} as

$$f_{inst} = \frac{1}{2\pi}[\omega_c + m x_{BB}(t) * h(t)], \tag{3.67}$$

suggesting that

$$\Delta f = \frac{1}{2\pi} m[x_{BB}(t) * h(t)]_{max}. \tag{3.68}$$

Since the peak voltage swing at the output of the Gaussian filter is approximately equal to that of $x_{BB}(t)$,

$$\frac{m}{2\pi} x_{BB,max} = 115 \text{ kHz}. \tag{3.69}$$

As explained in Chapter 8, m is a property of the voltage-controlled oscillator (called the "gain" of the VCO) and must be chosen to satisfy (3.69).

Bluetooth specifies an output level of 0 dBm (1 mW).[14] Along with the constant-envelope modulation, this relaxed value greatly simplifies the design of the power amplifier, allowing it to be a simple 50-Ω buffer.

The Bluetooth transmit spectrum mask is shown in Fig. 3.78. At an offset of 550 kHz from the center of the desired channel, the power measured in a 100-kHz bandwidth must be at least 20 dB below the TX power measured in the same bandwidth but in the center

14. This corresponds to the most common case of "power class 3." Other power classes with higher output levels have also been specified.

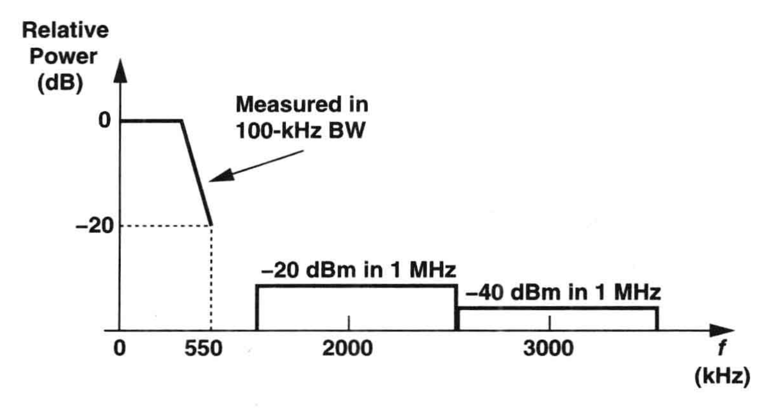


Figure 3.78 *Bluetooth transmission mask.*

of the channel. Moreover, the power measured in the entire 1-MHz alternate adjacent channel must remain below −20 dBm. Similarly, the power in the third and higher adjacent channels must be less than −40 dBm.[15]

A Bluetooth TX must minimally interfere with cellular and WLAN systems. For example, it must produce in a 100-kHz bandwidth less than −47 dBm in the range of 1.8 GHz to 1.9 GHz or 5.15 GHz to 5.3 GHz.

The carrier frequency of each Bluetooth carrier has a tolerance of ±75 kHz (≈ ±30 ppm). Since the carrier synthesis is based on a reference crystal frequency (Chapter 10), the crystal must have an error of less than ±30 ppm.

Receiver Characteristics Bluetooth operates with a reference sensitivity of −70 dBm (for $BER = 10^{-3}$), but most commercial products push this value to about −80 dBm, affording longer range.

Example 3.14

Estimate the NF required of a Bluetooth receiver.

Solution:

Assuming an SNR of 17 dB and a channel bandwidth of 1 MHz, we obtain a noise figure of 27 dB for a sensitivity of −70 dBm. It is this very relaxed NF that allows manufacturers to push for higher sensitivities (lower noise figures).

Figure 3.79 illustrates the blocking tests specified by Bluetooth. In Fig. 3.79(a), the desired signal is 10 dB higher than the reference sensitivity and another modulated Bluetooth signal is placed in the adjacent channel (with equal power) or in the alternate adjacent channel (with a power of −30 dBm. These specifications in turn require a sharp roll-off in the analog baseband filters (Chapter 4).

15. Up to three exceptions are allowed for the third and higher adjacent channel powers, with a relaxed specification of −20 dBm.

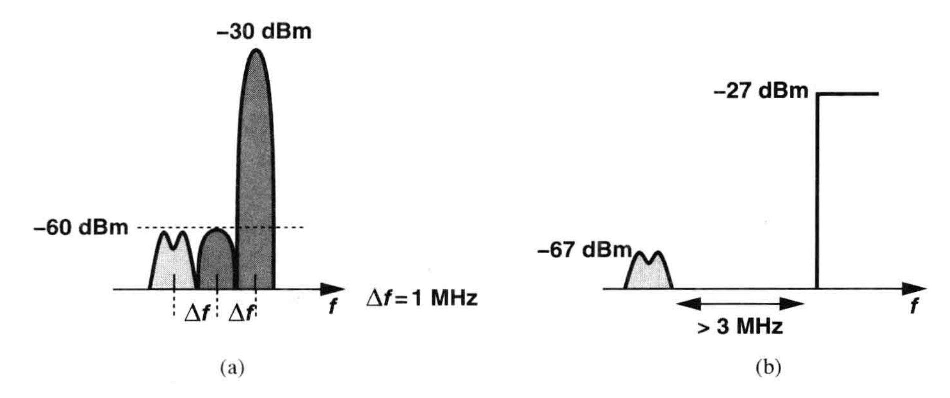


Figure 3.79 *Bluetooth receiver blocking test for (a) adjacent and alternate channels, and (b) channels at > 3-MHz offset.*

In Fig. 3.79(b), the desired signal is 3 dB above the sensitivity and a modulated blocker is applied in the third or higher adjacent channel with a power of −27 dBm. Thus, the 1-dB compression point of the receiver must exceed this value.

A Bluetooth receiver must also withstand out-of-band sinusoidal blockers. As shown in Fig. 3.80, with the desired signal at −67 dBm, a tone level of −27 dBm or −10 dBm must be tolerated according to the tone frequency range. We observe that if the receiver achieves a P_{1dB} of several dB above −27 dBm, then the filter following the antenna has a relaxed out-of-band attenuation requirement.

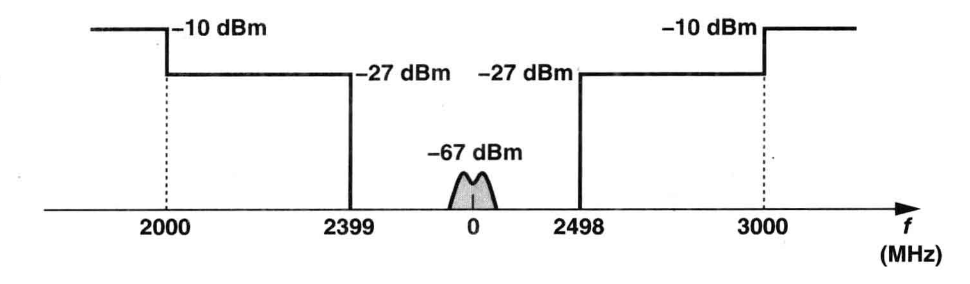


Figure 3.80 *Bluetooth receiver out-of-band blocking test.*

The intermodulation test in Bluetooth is depicted in Fig. 3.81. The desired signal level is 6 dB higher than the reference sensitivity and the blockers are applied at −39 dBm with $\Delta f = 3$, 4, or 5 MHz. In Problem 3.6, we derive the required RX IP_3 and note that it is quite relaxed.

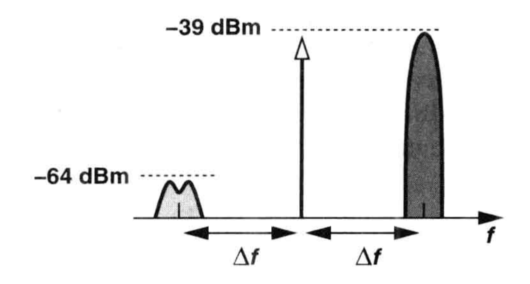


Figure 3.81 *Bluetooth receiver intermodulation test.*

Bluetooth also stipulates a *maximum* usable input level of −20 dBm. That is, a desired channel at this level must be detected properly (with $BER=10^{-3}$) by the receiver.

Example 3.15

Does the maximum usable input specification pose any design constraints?

Solution:

Yes, it does. Recall that the receiver must detect a signal as low as −60 dBm; i.e., the receiver chain must provide enough gain before detection. Suppose this gain is about 60 dB, yielding a signal level of around 0 dBm (632 mV_{pp}) at the end of the chain. Now, if the received signal rises to −20 dBm, the RX output must reach +40 dBm (63.2 V_{pp}), unless the chain becomes heavily nonlinear. The nonlinearity may appear benign as the signal has a constant envelope, but the heavy saturation of the stages may distort the baseband data. For this reason, the receiver must incorporate "automatic gain control" (AGC), reducing the gain of each stage as the input signal level increases (Chapter 13).

3.7.5 IEEE802.11a/b/g

The IEEE802.11a/b/g standard allows high-speed wireless connectivity, providing a maximum data rate of 54 Mb/s. The 11a and 11g versions are identical except for their frequency bands (5 GHz and 2.4 GHz, respectively). The 11b version also operates in the 2.4-GHz band but with different characteristics. The 11g and 11b standards are also known as "WiFi." We begin our study with 11a/g.

The 11a/g standard specifies a channel spacing of 20 MHz with different modulation schemes for different data rates. Figure 3.82 shows the air interface and channelization of 11a. We note that higher data rates use denser modulation schemes, posing tougher demands on the TX and RX design. Also, as explained in Section 3.3.6, for rates higher than a few megabits per second, wireless systems employ OFDM so as to minimize the effect of delay spread. This standard incorporates a total of 52 subcarriers with a spacing of 0.3125 MHz (Fig. 3.83). The middle subchannel and the first and last five subchannels are unused. Moreover, four of the subcarriers are occupied by BPSK-modulated "pilots" to simplify the detection in the receiver in the presence of frequency offsets and phase noise. Each OFDM symbol is 4 μs long.

The TX must deliver a power of at least 40 mW (+16 dBm) while complying with the spectrum mask shown in Fig. 3.84. Here, each point represents the power measured in a 100-kHz bandwidth normalized to the overall output power. The sharp drop between 9 MHz and 11 MHz calls for pulse shaping in the TX baseband (Section 3.3.1). In fact, pulse shaping reduces the channel bandwidth to 16.6 MHz. The carrier frequency has a tolerance of ±20 ppm. Also, the carrier leakage must remain 15 dB below the overall output power.

The receiver sensitivity in 11a/g is specified in conjunction with the data rate. Table 3.1 summarizes the sensitivities along with adjacent channel and alternate adjacent channel levels. The "packet error rate" must not exceed 10%, corresponding to a BER of less than 10^{-5}.

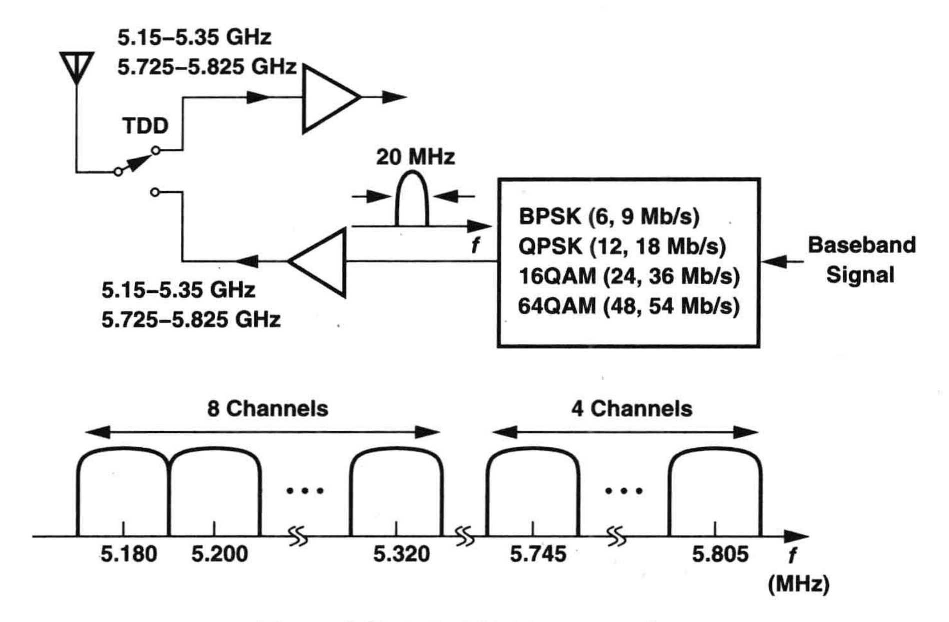


Figure 3.82 *IEEE802.11a air interface.*

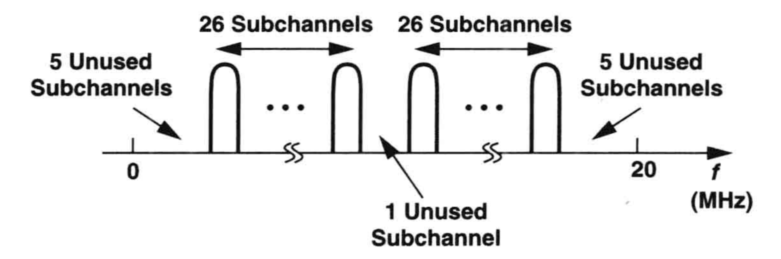


Figure 3.83 *OFDM channelization in 11a.*

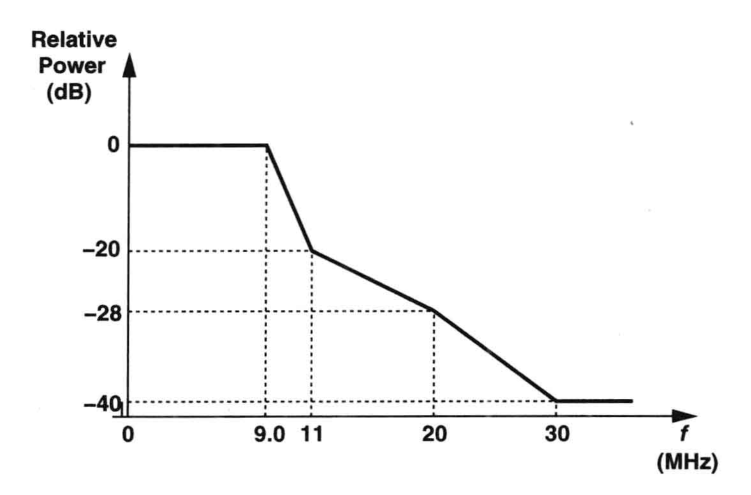


Figure 3.84 *IEEE802.11a transmission mask.*

Table 3.1 *IEEE802.11a data rates, sensitivities, and adjacent channel levels.*

Data Rate (Mb/s)	Reference Sensitivity (dBm)	Adj. Channel Level (dB)	Alt. Channel Level (dB)
6.0	−82	16	32
9.0	−81	15	31
12	−79	13	29
18	−77	11	27
24	−74	8.0	24
36	−70	4.0	20
48	−66	0	16
54	−65	−1	15

Example 3.16

Estimate the noise figure necessary for 6-Mb/s and 54-Mb/s reception in 11a/g.

Solution:

First, consider the rate of 6 Mb/s. Assuming a noise bandwidth of 20 MHz, we obtain 19 dB for the sum of the NF and the required SNR. Similarly, for the rate of 54 Mb/s, this sum reaches 36 dB. An NF of 10 dB leaves an SNR of 9 dB for BPSK and 26 dB for 64QAM, both sufficient for the required error rate. In fact, most commercial products target an NF of about 6 dB so as to achieve a sensitivity of about −70 dBm at the highest date rate.

The large difference between the sensitivities in Table 3.1 does make the receiver design difficult: the gain of the chain must reach about 82 dB in the low-rate case and be reduced to about 65 dB in the high-rate case.[16]

The adjacent channel tests shown in Table 3.1 are carried out with the desired channel at 3 dB above the reference sensitivity and another modulated signal in the adjacent or alternate channel.

An 11a/g receiver must operate properly with a maximum input level of −30 dBm. As explained for Bluetooth in Section 3.7.4, such a high input amplitude saturates the receiver chain, a very serious issue for the denser modulations used in 11a/g. Thus, the RX gain must be programmable from about 82 dB to around 30 dB.

16. As a rule of thumb, a receiver analog baseband output should be around 0 dBm.

Example 3.17

Estimate the 1-dB compression point necessary for 11a/g receivers.

Solution:

With an input of −30 dBm, the receiver must not compress. Furthermore, recall from Section 3.3.6 that an OFDM signal having N subchannels exhibits a peak-to-average ratio of about $2 \ln N$. For $N = 52$, we have $PAR = 7.9$. Thus, the receiver must not compress even for an input level reaching −30 dBm + 7.9 dB = −22.1 dBm. The envelope variation due to baseband pulse shaping may require an even higher P_{1dB}.

The IEEE802.11b supports a maximum data rate of 11Mb/s with "complementary code keying" (CCK) modulation.[17] But under high signal loss conditions, the data rate is scaled down to 5.5 Mb/s, 2 Mb/s, or 1 Mb/s. The last two rates employ QPSK and BPSK modulation, respectively. Each channel of 11b occupies 22 MHz in the 2.4-GHz ISM band. To offer greater flexibility, 11b specifies *overlapping* channel frequencies (Fig. 3.85). Of course, users operating in close proximity of one another avoid overlapping channels. The carrier frequency tolerance is ±25 ppm.

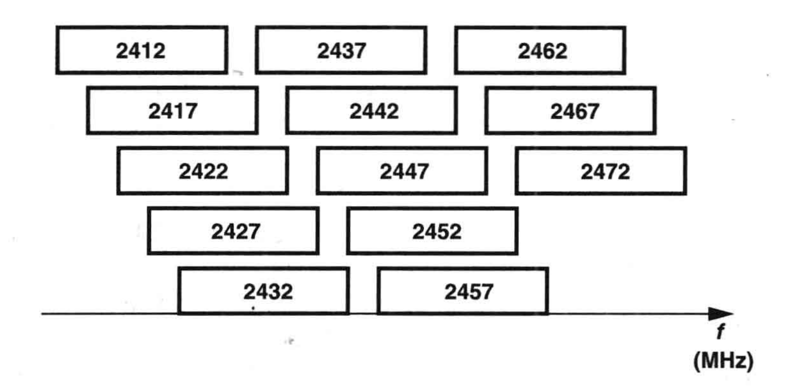


Figure 3.85 *Overlapping channelization in 11a.*

The 11b standard stipulates a TX output power of 100 mW (+20 dBm) with the spectrum mask shown in Fig. 3.86, where each point denotes the power measured in a 100-kHz bandwidth. The low emission in adjacent channels dictates the use of pulse shaping in the TX baseband. The standard also requires that the carrier leakage be 15 dB below the peak of the spectrum in Fig. 3.86.[18]

An 11b receiver must achieve a sensitivity of −76 dBm for a "frame error rate" of 8×10^{-2} and operate with input levels as high as −10 dBm. The adjacent channel can be 35 dB above the desired signal, with the latter at −70 dBm.

17. CCK is a variant of QPSK.

18. Note that, unlike the 11a/g specification, this leakage is *not* with respect to the overall TX output power.

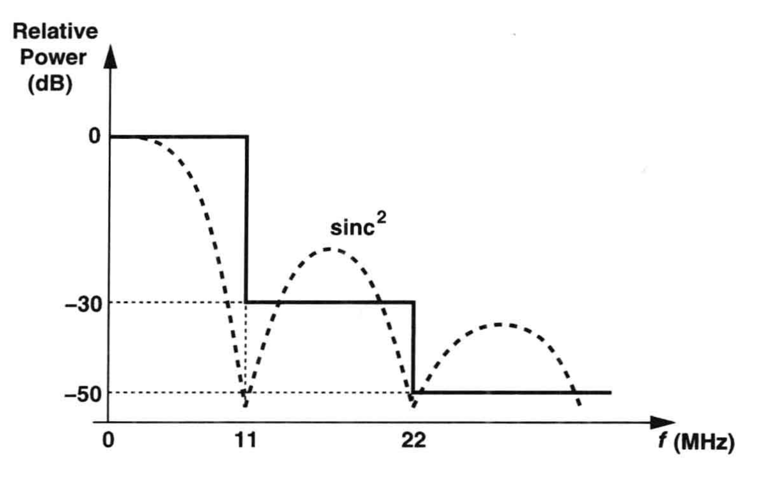


Figure 3.86 *IEEE802.11b transmission mask.*

3.8 APPENDIX I: DIFFERENTIAL PHASE SHIFT KEYING

A difficulty in the detection of PSK signals is that the phase relates to the time origin and has no "absolute" meaning. For example, a 90° phase shift in a QPSK waveform converts the constellation to a similar one, but with all the symbols interpreted incorrectly. Thus, simple PSK waveforms cannot be detected noncoherently. However, if the information lies in the phase *change* from one bit (or symbol) to the next, then a time origin is not required and noncoherent detection is possible. This is accomplished through "differential" encoding and decoding of the baseband signal before modulation and after demodulation, respectively.

Let us consider binary differential PSK (DPSK). The rule for differential encoding is that if the present input bit is a ONE, then the output state of the encoder does not

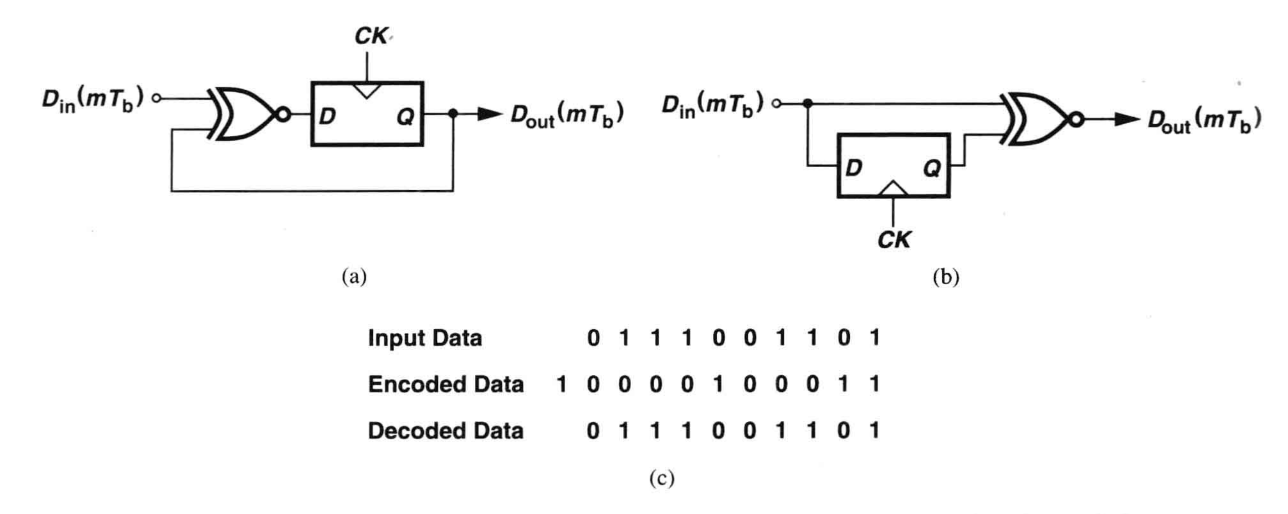


Figure 3.87 *(a) Differential encoding, (b) differential decoding, (c) example of encoded and decoded sequence.*

change, and vice versa. This requires an extra starting bit (of arbitrary value). The concept can be better understood by considering the implementation depicted in Fig. 3.87(a). An exclusive-NOR (XNOR) gate compares the present output bit, $D_{out}(mT_b)$, with the present input bit, $D_{in}(mT_b)$, to determine the next output state:

$$D_{out}[(m+1)T_b] = \overline{D_{in}(mT_b) \oplus D_{out}(mT_b)}, \tag{3.70}$$

implying that if $D_{in}(mT_b) = 1$, then $D_{out}[(m+1)T_b] = D_{out}(mT_b)$, and if $D_{in}(mT_b) = 0$ then $D_{out}[(m+1)T_b] = \overline{D_{out}(mT_b)}$. The extra starting bit mentioned above corresponds to the state of the flipflop before the data sequence begins.

REFERENCES

[1] L. W. Couch, *Digital and Analog Communication Systems,* Fourth Edition, New York: Macmillan Co., 1993.

[2] H. E. Rowe, *Signals and Noise in Communication Systems,* New Jersey: Van Nostrand Co., 1965.

[3] R. E. Ziemer and R. L. Peterson, *Digital Communication and Spread Spectrum Systems,* New York: Macmillan, 1985.

[4] P. A. Baker, "Phase-Modulated Data Sets for Serial Transmission at 2000 and 2400 Bits per Second," Part I, *AIEE Trans. Communication Electronics,* pp. 166–171, July 1962.

[5] Y. Akaiwa and Y. Nagata, "Highly Efficient Digital Mobile Communication with a Linear Modulation Method," *IEEE J. of Selected Areas in Commnunications,* vol. 5, pp. 890–895, June 1987.

[6] N. Dinur and D. Wulich, "Peak-to-average power ratio in high-order OFDM," *IEEE Tran. Comm.,* vol. 49, pp. 1063–1072, June 2001.

[7] T. S. Rappaport, *Wireless Communications, Principles and Practice,* New Jersey: Prentice Hall, 1996.

[8] D. P. Whipple, "North American Cellular CDMA," *Hewlett-Packard Journal,* pp. 90–97, Dec. 1993.

[9] A. Salmasi and K. S. Gilhousen, "On the System Design Aspects of Code Division Multiple Access (CDMA) Applied to Digital Cellular and Personal Communications Networks," *Proc. IEEE Veh. Tech. Conf.,* pp. 57–62, May 1991.

[10] R. Kerr et al., "The CDMA Digitial Cellular System, An ASIC Overview," *Proceedings of IEEE CICC,* pp. 10.1.1–10.1.7, May 1992.

[11] J. Hinderling et al., "CDMA Mobile Station Modem ASIC," *Proceedings of IEEE CICC,* pp. 10.2.1–10.2.5, May 1992.

PROBLEMS

3.1. Due to imperfections, a 16QAM generator produces $\alpha_1 A_c \cos(\omega_c t + \Delta\theta) - \alpha_2 A_c (1+\epsilon) \sin\omega_c t$, where $\alpha_1 = \pm 1, \pm 2$ and $\alpha_2 = \pm 1, \pm 2$.
(a) Construct the signal constellation for $\Delta\theta \neq 0$ but $\epsilon = 0$.
(b) Construct the signal constellation for $\Delta\theta = 0$ but $\epsilon \neq 0$.

3.2. Repeat Example 3.12 if the noise figure is less than 10 dB.

3.3. Repeat Example 3.12 for WCDMA.

3.4. Determine the maximum tolerable relative noise floor (in dBc/Hz) that an IMT-2000 TX can generate in the DCS1800 band.

3.5. Repeat Example 3.12 for the scenario shown in Fig. 3.73.

3.6. From Fig. 3.81, estimate the required IP_3 of a Bluetooth receiver.

3.7. A "ternary" FSK signal can be defined as

$$x_{FSK}(t) = a_1 \cos \omega_1 t + a_2 \cos \omega_2 t + a_3 \cos \omega_3 t, \tag{3.71}$$

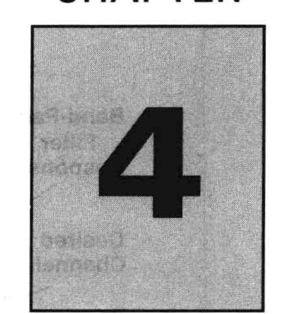
where only one of the coefficients is equal to 1 at a time and the other two are equal to zero. Plot the constellation of this signal.

3.8. In order to detect (demodulate) an AM signal, we can multiply it by the LO waveform and apply the result to a low-pass filter. Beginning with Eq. (3.2), explain the operation of the detector.

3.9. Repeat the above problem for the BPSK signal expressed by Eq. (3.26).

3.10. The BPSK signal expressed by Eq. (3.26) is to be demodulated. As studied in the previous problem, we must multiply $a_n \cos \omega_c t$ by an LO waveform. Now suppose the LO waveform generated in a receiver has a slight "frequency offset" with respect to the incoming carrier. That is, we in fact multiply $a_n \cos \omega_c t$ by $\cos(\omega_c + \Delta\omega)t$. Prove that the signal constellation *rotates* with time at a rate of $\Delta\omega$.

CHAPTER



TRANSCEIVER ARCHITECTURES

With the understanding developed in previous chapters of RF design and communication principles, we are now prepared to move down to the transceiver architecture level. The choice of an architecture is determined by not only the RF performance that it can provide but other parameters such as complexity, cost, power dissipation, and the number of external components. In the past ten years, it has become clear that high levels of integration improve the system performance along all of these axes. It has also become clear that architecture design and circuit design are inextricably linked, requiring iterations between the two. The outline of the chapter is shown below.

Heterodyne Receivers	Direct-Conversion Receivers	Image-Reject and Low-IF Receivers	Transmitter Architectures
■ Problem of Image	■ LO Leakage and Offsets	■ Hartley and Weaver Receivers	■ TX Baseband Processing
■ Mixing Spurs	■ Even-Order Nonlinearity	■ Low-IF Receivers	■ Direct-Conversion TX
■ Sliding-IF RX	■ I/Q Mismatch	■ Polyphase Filters	■ Heterodyne and Sliding-IF TX

4.1 GENERAL CONSIDERATIONS

The wireless communications environment is often called "hostile" to emphasize the severe constraints that it imposes on transceiver design. Perhaps the most important constraint originates from the limited channel bandwidth allocated to each user (e.g., 200 kHz in GSM). From Shannon's theorem,[1] this translates to a limited rate of information, dictating the use of sophisticated baseband processing techniques such as coding, compression, and bandwidth-efficient modulation.

1. Shannon's theorem states that the achievable data rate of a communication channel is equal to $B\log_2(1 + SNR)$, where B denotes the bandwidth and SNR the signal-to-noise ratio.

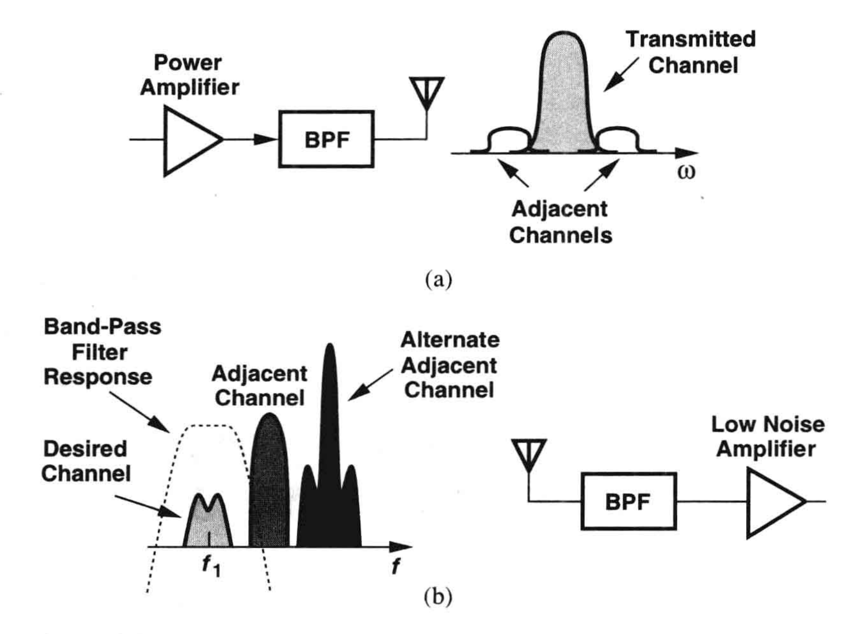


Figure 4.1 *(a) Transmitter and (b) receiver front ends of a wireless system.*

The narrow channel bandwidth also impacts the RF design of the transceiver. As depicted in Fig. 4.1, the transmitter must employ narrowband modulation and amplification to avoid leakage to adjacent channels, and the receiver must be able to process the desired channel while sufficiently rejecting strong in-band and out-of-band interferers.

The reader may recall from Chapter 2 that both nonlinearity and noise increase as we add more stages to a cascade. In particular, we recognized that the linearity of a receiver must be high enough to accommodate interferers without experiencing compression or significant intermodulation. The reader may then wonder if we can simply filter the interferers so as to relax the receiver linearity requirements. Unfortunately, two issues arise here. First, since an interferer may fall only one or two channels away from the desired signal (Fig. 4.2), the filter must provide a very high selectivity (i.e., a high Q). If the interferer level is, say, 50–60 dB above the desired signal level, the required value of Q reaches prohibitively high values, e.g., millions. Second, since a different carrier frequency may be allocated to the

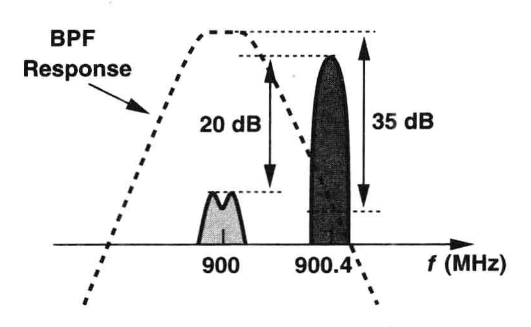


Figure 4.2 *Hypothetical filter to suppress an interferer.*

user at different times, such a filter would need a *variable*, yet precise, center frequency—a property very difficult to implement.

Example 4.1

A 900-MHz GSM receiver with 200-kHz channel spacing must tolerate an alternate adjacent channel blocker 20 dB higher than the desired signal. Calculate the Q of a second-order LC filter required to suppress this interferer by 35 dB.

Solution:

As shown in Fig. 4.2, the filter frequency response must provide an attenuation of 35 dB at 400 kHz away from center frequency of 900 MHz. For a second-order RLC tank, we write the impedance as

$$Z_T(s) = \frac{RLs}{RLCs^2 + Ls + R}, \tag{4.1}$$

and assume a resonance frequency of $\omega_0 = 1/\sqrt{LC} = 2\pi(900 \text{ MHz})$. The magnitude squared of the impedance is thus given by

$$|Z_T(j\omega)|^2 = \frac{L^2\omega^2}{(1 - LC\omega^2)^2 + L^2\omega^2/R^2}. \tag{4.2}$$

For an attenuation of 35 dB (=56.2) at 900.4 MHz, this quantity must be equal to $R^2/56.2^2$ (why?). Solving for $L^2\omega^2/R^2$, we obtain

$$\frac{L^2\omega^2}{R^2} = 2.504 \times 10^{-10}. \tag{4.3}$$

Recall from Chapter 2 that $Q = R/(L\omega) = 63{,}200$.

Channel Selection and Band Selection The type of filtering speculated above is called "channel-selection filtering" to indicate that it "selects" the desired signal channel and "rejects" the interferers in the other channels. We make two key observations here: (1) all of the stages in the receiver chain that *precede* channel-selection filtering must be sufficiently linear to avoid compression or excessive intermodulation, and (2) since channel-selection filtering is extremely difficult at the input carrier frequency, it must be deferred to some other point along the chain where the center frequency of the desired channel is substantially *lower* and hence the required filter Q's are more reasonable.[2]

Nonetheless, most receiver front ends do incorporate a "band-select" filter, which selects the entire *receive band* and rejects "out-of-band" interferers (Fig. 4.3), thereby suppressing components that may be generated by users that do not belong to the standard

2. The Q of a band-pass filter may be roughly defined as the center frequency divided by the −3-dB bandwidth.

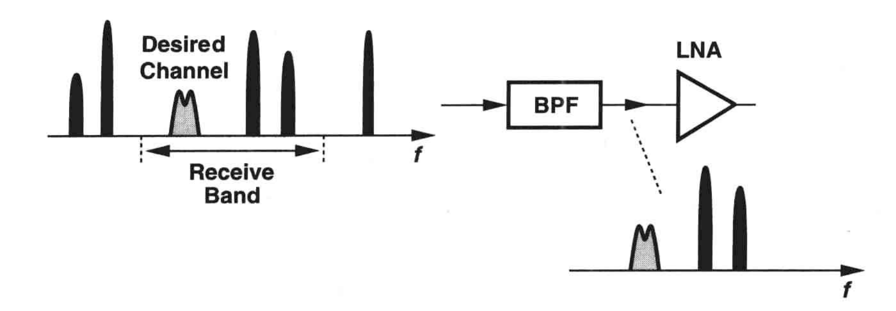


Figure 4.3 *Band-selection filtering.*

of interest. We thus distinguish between out-of-band interferers and "in-band interferers," which are typically removed near the end of the receiver chain.

The front-end band-select (passive) filter suffers from a trade-off between its selectivity and its in-band loss because the edges of the band-pass frequency response can be sharpened only by increasing the order of the filter, i.e., the number of cascaded sections within the filter. Now we note from Chapter 2 that the front-end loss directly raises the NF of the entire receiver, proving very undesirable. The filter is therefore designed to exhibit a small loss (0.5 to 1 dB) and some frequency selectivity.

Figure 4.4 plots the frequency response of a typical duplexer,[3] exhibiting an in-band loss of about 2 dB and an out-of-band rejection of 30 dB at 20-MHz "offset" with respect to the receive band. That is, an interferer appearing at f_1 (20 MHz away from the RX band) is attenuated by only 30 dB, a critical issue in the design of both the receive path and the frequency synthesizer (Chapter 10).

The in-band loss of the above duplexer in the *transmit* band also proves problematic as it "wastes" some of the power amplifier output. For example, with 2-dB of loss and a 1-W

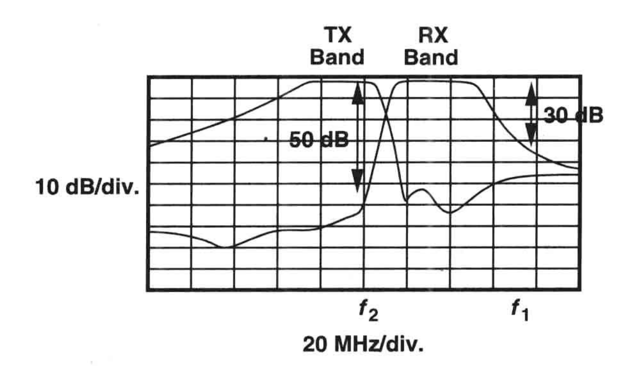


Figure 4.4 *Duplexer characteristics.*

3. As mentioned in Chapter 3, a duplexer consists of two band-pass filters, one for the TX band and another for the RX band.

PA, as much as 370 mW is dissipated within the duplexer—more than the typical power consumed by the entire receive path!

Our observations also indicate the importance of controlled spectral regrowth through proper choice of the modulation scheme and the power amplifier (Chapter 3). The out-of-channel energy produced by the PA *cannot* be suppressed by the front-end BPF and must be acceptably small by design.

TX-RX Feedthrough As mentioned in Chapter 3, TDD systems activate only the TX or the RX at any point in time to avoid coupling between the two. Also, even though an FDD system, GSM offsets the TX and RX time slots for the same reason. On the other hand, in full-duplex standards, the TX and the RX operate concurrently. (As explained in Chapter 3, CDMA systems require continual power control and hence concurrent TX and RX operation.) We recognize from the typical duplexer characteristics shown in Fig. 4.4 that the transmitter output at frequencies near the upper end of the TX band, e.g., at f_2, is attenuated by only about 50 dB as it leaks to the receiver. Thus, with a 1-W TX power, the leakage sensed by the LNA can reach -20 dBm (Fig. 4.5), dictating a substantially higher RX compression point. For this reason, CDMA receivers must meet difficult linearity requirements.

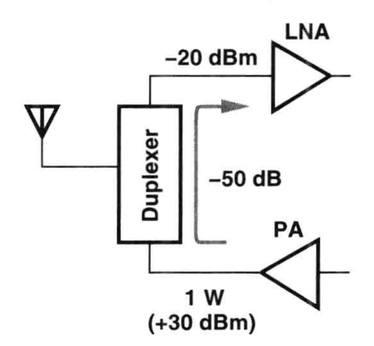


Figure 4.5 *TX leakage in a CDMA transceiver.*

Example 4.2

Explain how a band-pass filter following the LNA can alleviate the TX-RX leakage in a CDMA system.

Solution:

As depicted in Fig. 4.6, if the BPF provides additional rejection in the TX band, the linearity required of the rest of the RX chain is proportionally relaxed. The LNA compression point, however, must still be sufficiently high.

(Continues)

Example 4.2 *(Continued)*

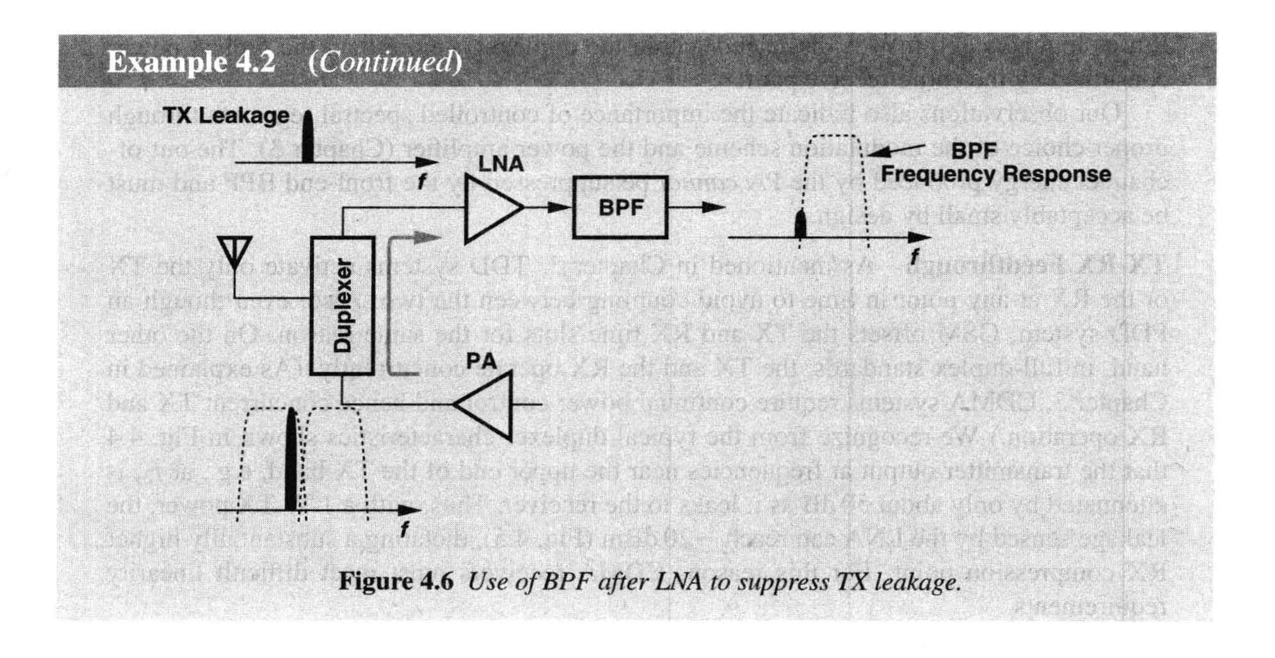


Figure 4.6 *Use of BPF after LNA to suppress TX leakage.*

4.2 RECEIVER ARCHITECTURES

4.2.1 Basic Heterodyne Receivers

As mentioned above, channel-selection filtering proves very difficult at high carrier frequencies. We must devise a method of "translating" the desired channel to a much lower center frequency so as to permit channel-selection filtering with a reasonable Q. Illustrated in Fig. 4.7(a), the translation is performed by means of a "mixer," which in this chapter is viewed as a simple analog multiplier. To lower the center frequency, the signal is multiplied by a sinusoid $A_0 \cos \omega_{LO} t$, which is generated by a local oscillator (LO). Since multiplication in the time domain corresponds to convolution in the frequency domain, we observe from Fig. 4.7(b) that the impulses at $\pm\omega_{LO}$ shift the desired channel to $\pm(\omega_{in} \pm \omega_{LO})$. The components at $\pm(\omega_{in} + \omega_{LO})$ are not of interest and are removed by the low-pass filter (LPF) in Fig. 4.7(a), leaving the signal at a center frequency of $\omega_{in} - \omega_{LO}$. This operation is called "downconversion mixing" or simply "downconversion." Due to its high noise, the downconversion mixer is preceded by a low-noise amplifier [Fig. 4.7(c)].

Called the intermediate frequency (IF), the center of the downconverted channel, $\omega_{in} - \omega_{LO}$, plays a critical role in the performance. "Heterodyne" receivers employ an LO frequency unequal to ω_{in} and hence a nonzero IF.[4]

How does a heterodyne receiver cover a given frequency band? For an N-channel band, we can envision two possibilities. (1) The LO frequency is *constant* and each RF channel

4. In this book, we do not distinguish between heterodyne and "super heterodyne" architectures. The term *heterodyne* derives from *hetero* (different) and *dyne* (to mix).

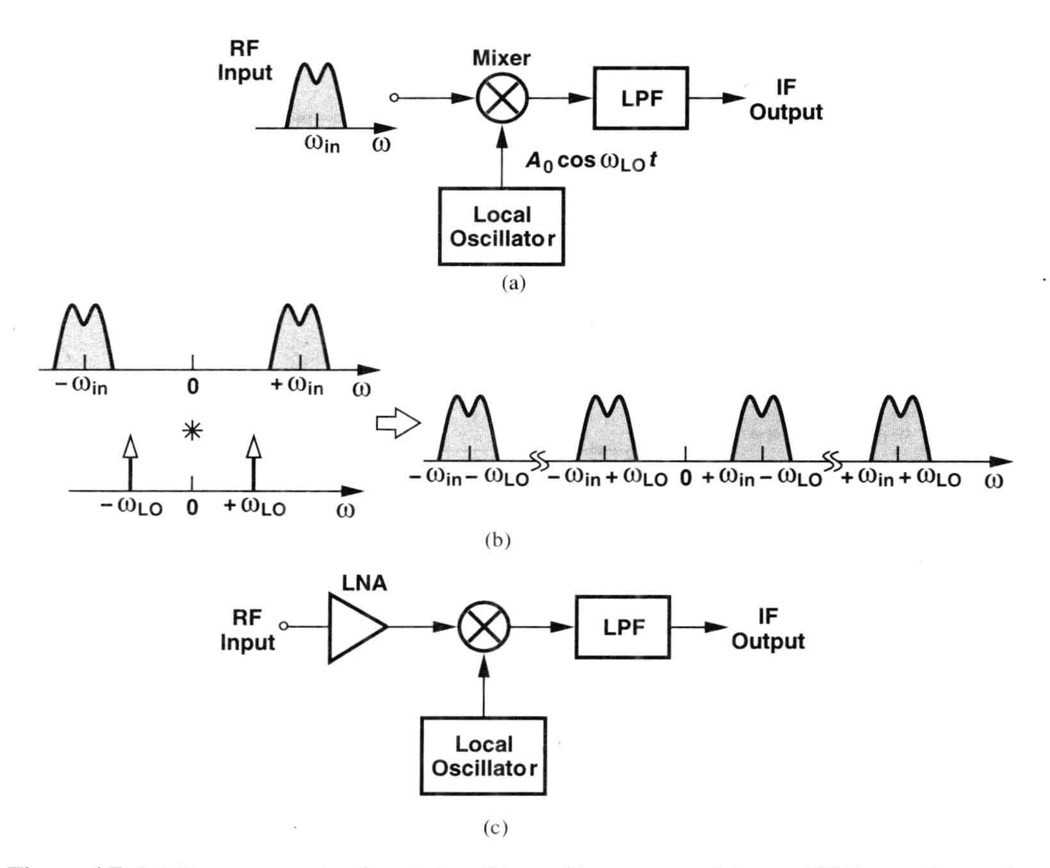


Figure 4.7 *(a) Downconversion by mixing, (b) resulting spectra, (c) use of LNA to reduce noise.*

is downconverted to a different IF channel [Fig. 4.8(a)], i.e., $f_{IFj} = f_{RFj} - f_{LO}$. (2) The LO frequency is *variable* so that all RF channels within the band of interest are translated to a single value of IF [Fig. 4.8(b)], i.e., $f_{LOj} = f_{RFj} - f_{IF}$. The latter case is more common as it simplifies the design of the IF path; e.g., it does not require a filter with a variable center frequency to select the IF channel of interest and reject the others. However, this approach demands a feedback loop that precisely defines the LO frequency steps, i.e., a "frequency synthesizer" (Chapters 9–11).

Problem of Image Heterodyne receivers suffer from an effect called the "image." To understand this phenomenon, let us assume a sinusoidal input and express the IF component as

$$A \cos \omega_{IF} t = A \cos(\omega_{in} - \omega_{LO})t \tag{4.4}$$

$$= A \cos(\omega_{LO} - \omega_{in})t. \tag{4.5}$$

That is, whether $\omega_{in} - \omega_{LO}$ is positive or negative, it yields the same intermediate frequency. Thus, whether ω_{in} lies *above* ω_{LO} or *below* ω_{LO}, it is translated to the same IF. Figure 4.9

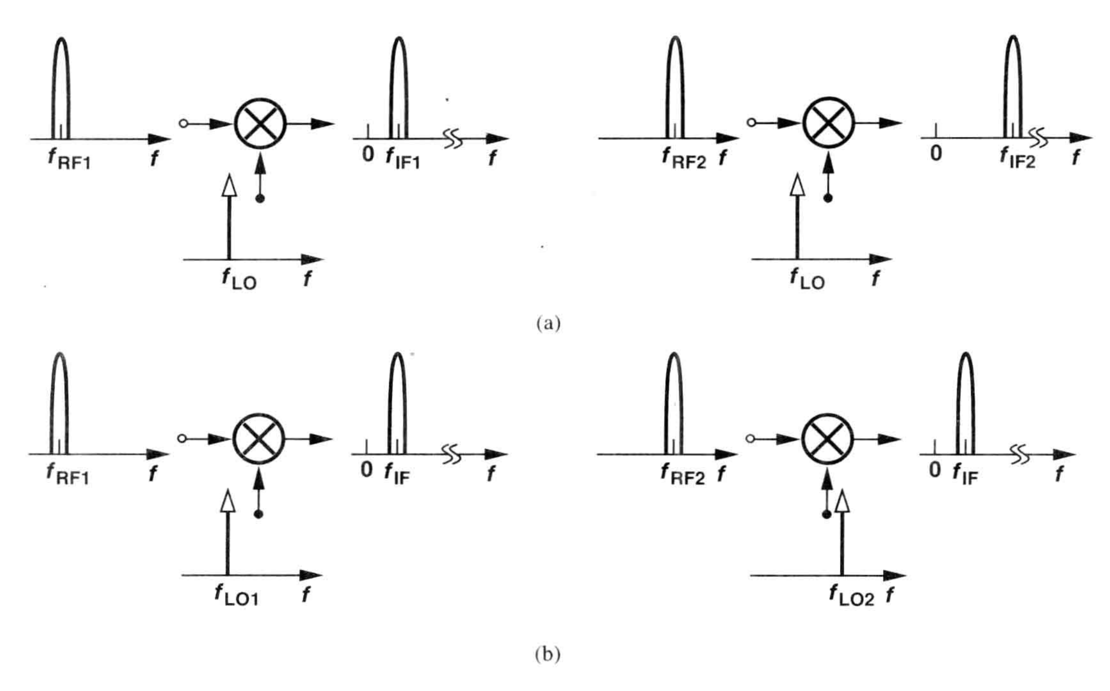


Figure 4.8 *(a) Constant-LO and (b) constant-IF downconversion mixing.*

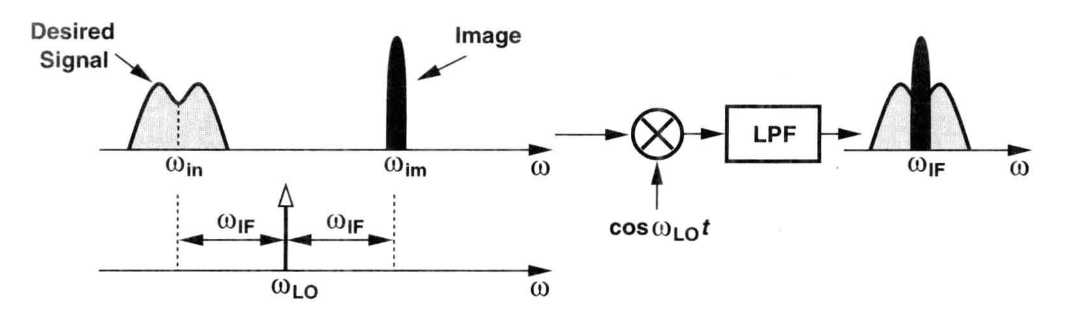


Figure 4.9 *Problem of image in heterodyne downconversion.*

depicts a more general case, revealing that two spectra located *symmetrically* around ω_{LO} are downconverted to the IF. Due to this symmetry, the component at ω_{im} is called the image of the desired signal. Note that $\omega_{im} = \omega_{in} + 2\omega_{IF} = 2\omega_{LO} - \omega_{in}$.

What creates the image? The numerous users in all standards (from police to WLAN bands) that transmit signals produce many interferers. If one interferer happens to fall at $\omega_{im} = 2\omega_{LO} - \omega_{in}$, then it corrupts the desired signal after downconversion.

While each wireless standard imposes constraints upon the emissions by its own users, it may have no control over the signals in other bands. The image power can therefore be much higher than that of the desired signal, requiring proper "image rejection."

Example 4.3

Suppose two channels at ω_1 and ω_2 have been received and $\omega_1 < \omega_2$. Study the downconverted spectrum as the LO frequency varies from below ω_1 to above ω_2.

Solution:

Shown in Fig. 4.10(a) is the case of $\omega_{LO} < \omega_1$. We note that the impulse at $-\omega_{LO}$ shifts the components at $+\omega_1$ and $+\omega_2$ to the left. Similarly, the impulse at $+\omega_{LO}$ shifts

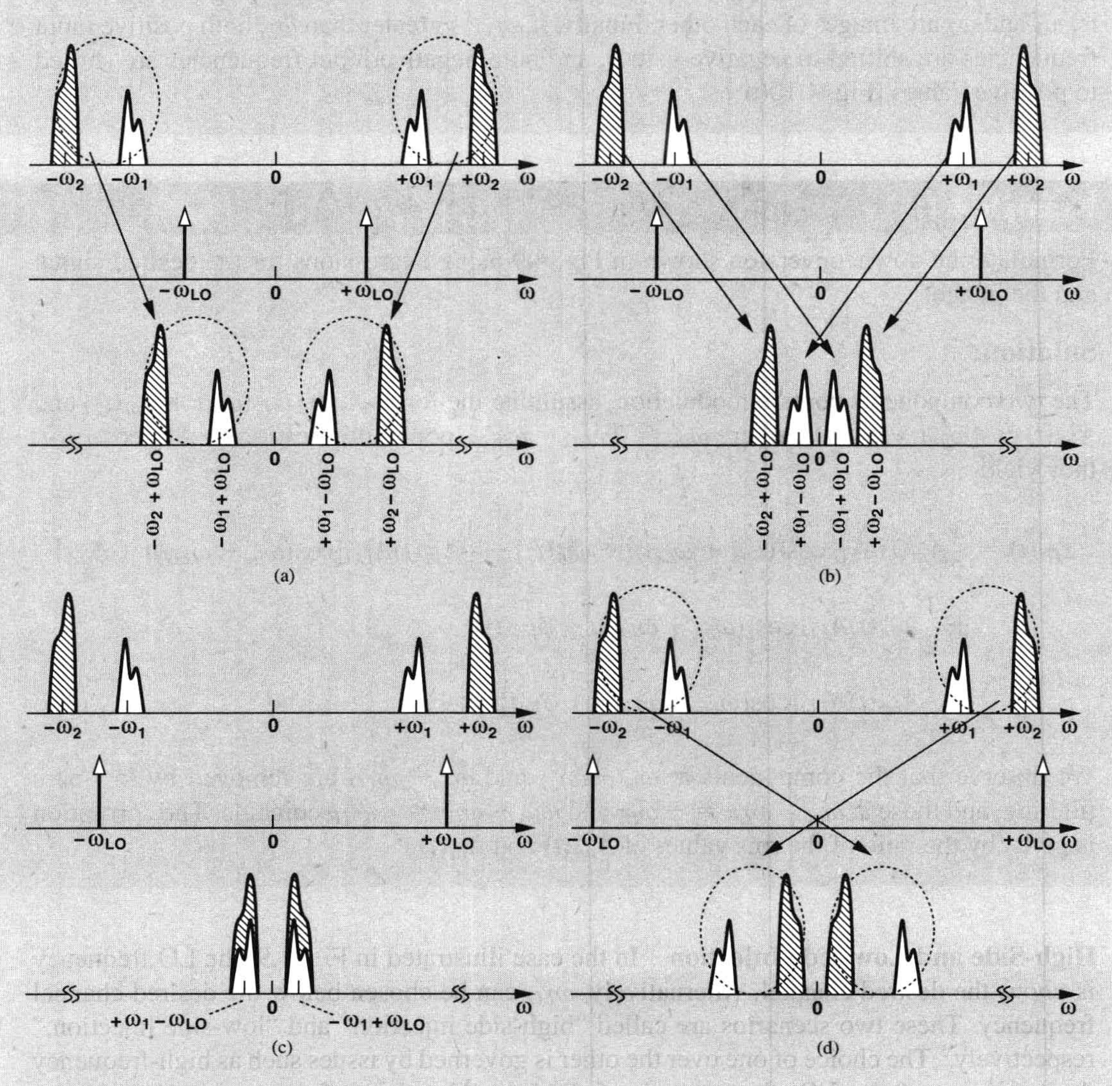


Figure 4.10 *Downconversion of two channels for (a) $\omega_{LO} < \omega_1$, (b) ω_{LO} slightly above ω_1, (c) ω_{LO} midway between ω_1 and ω_2, and (d) $\omega_{LO} > \omega_2$.*

(Continues)

Example 4.3 *(Continued)*

the components at $-\omega_1$ and $-\omega_2$ to the right. Since $\omega_{LO} < \omega_1$, the positive input frequencies remain positive after downconversion, and the negative input frequencies remain negative.

Now consider the case depicted in Fig. 4.10(b), where ω_{LO} is slightly greater than ω_1. Here, after downconversion the channel at $+\omega_1$ slides to *negative* frequencies while that at $+\omega_2$ remains positive. If ω_{LO} reaches $(\omega_1 + \omega_2)/2$, then the received channels are translated such that they completely overlap each other at the IF output [Fig. 4.10(c)]. That is, ω_1 and ω_2 are images of each other. Finally, if ω_{LO} is greater than ω_2, both positive input frequencies are shifted to negative values, and both negative input frequencies are shifted to positive values [Fig. 4.10(d)].

Example 4.4

Formulate the downconversion shown in Fig. 4.9 using expressions for the desired signal and the image.

Solution:

The two components contain modulation, assuming the forms $A_{in}(t)\cos[\omega_{in}t + \phi_{in}(t)]$ and $A_{im}(t)\cos[\omega_{im}t + \phi_{im}(t)]$, where $\omega_{im} = 2\omega_{LO} - \omega_{in}$. Upon multiplication by $A_{LO}\cos\omega_{LO}t$, they yield

$$x_{IF}(t) = \frac{1}{2}A_{in}(t)A_{LO}\cos[(\omega_{in} + \omega_{LO})t + \phi_{in}(t)] - \frac{1}{2}A_{in}(t)A_{LO}[\cos(\omega_{in} - \omega_{LO})t + \phi_{in}t]$$
$$+ \frac{1}{2}A_{im}(t)A_{LO}\cos[(\omega_{im} + \omega_{LO})t + \phi_{im}(t)]$$
$$- \frac{1}{2}A_{im}(t)A_{LO}[\cos(\omega_{im} - \omega_{LO})t + \phi_{im}t]. \tag{4.6}$$

We observe that the components at $\omega_{in} + \omega_{LO}$ and $\omega_{im} + \omega_{LO}$ are removed by low-pass filtering, and those at $\omega_{in} - \omega_{LO} = -\omega_{IF}$ and $\omega_{im} - \omega_{LO} = +\omega_{IF}$ coincide. The corruption is given by the ratio of the rms values of $A_{im}(t)$ and $A_{in}(t)$.

High-Side and Low-Side Injection In the case illustrated in Fig. 4.9, the LO frequency is *above* the desired channel. Alternatively, ω_{LO} can be chosen below the desired channel frequency. These two scenarios are called "high-side injection" and "low-side injection," respectively.[5] The choice of one over the other is governed by issues such as high-frequency design issues of the LO, the strength of the image-band interferers, and other system requirements.

5. These have also been called "superdyne" and "infradyne," respectively.

Example 4.5

The designer of an IEEE802.11g receiver attempts to place the image frequency in the GPS band, which contains only low-level satellite transmissions and hence no strong interferers. Is this possible?

Solution:

The two bands are shown in Fig. 4.11. The LO frequency must cover a range of 80 MHz but, unfortunately, the GPS band spans only 20 MHz. For example, if the lowest LO frequency is chosen so as to make 1.565 GHz the image of 2.4 GHz, then 802.11g channels above 2.42 GHz have images beyond the GPS band.

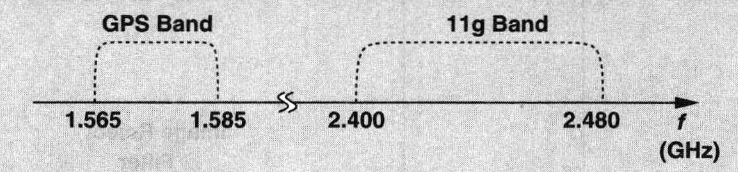


Figure 4.11 *Attempt to make the GPS band the image of an 11g receiver.*

Example 4.6

A dual-mode receiver is designed for both 802.11g and 802.11a. Can this receiver operate with a single LO?

Solution:

Figure 4.12(a) depicts the two bands. We choose the LO frequency halfway between the two so that a single LO covers the 11g band by high-side injection and the 11a band by low-side injection [Fig. 4.12(b)]. This greatly simplifies the design of the system but makes

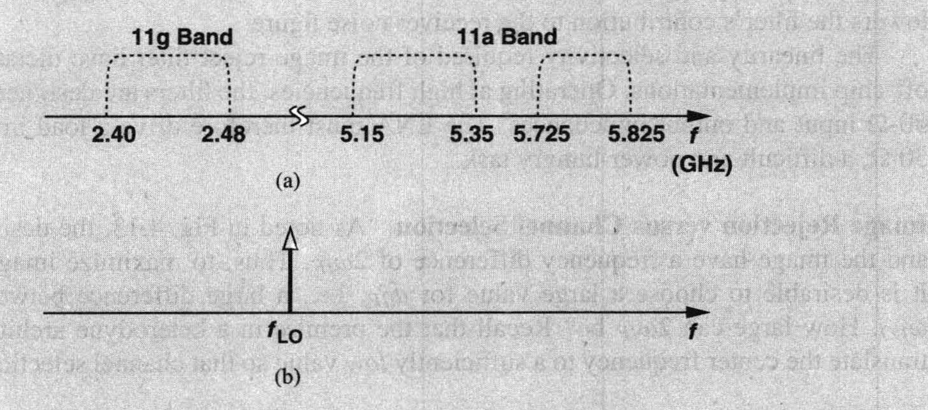


Figure 4.12 *(a) 11g and 11a bands, (b) choice of f_{LO}.*

(Continues)

Example 4.6 *(Continued)*

each band the image of the other. For example, if the receiver is in the 11a mode while an 11g transmitter operates in close proximity, the reception may be heavily corrupted. Note that also the IF in this case is quite high, an issue revisited later.

Image Rejection If the choice of the LO frequency leads to an image frequency in a high-interference band, the receiver must incorporate a means of suppressing the image. The most common approach is to precede the mixer with an "image-reject filter." As shown in Fig. 4.13, the filter exhibits a relatively small loss in the desired band and a large attenuation in the image band, two requirements that can be simultaneously met if $2\omega_{IF}$ is sufficiently large.

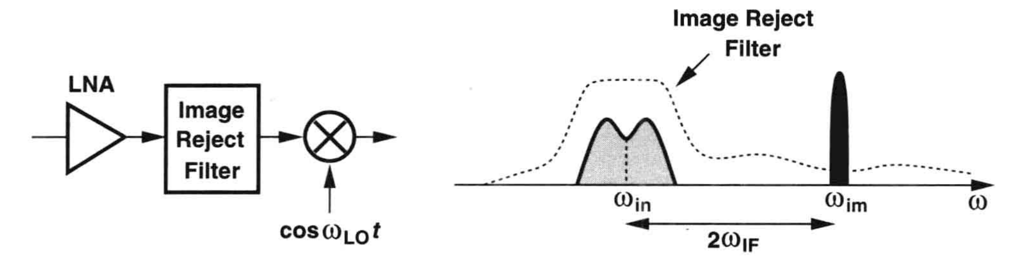


Figure 4.13 *Image rejection by filtering.*

Can the filter be placed before the LNA? More generally, can the front-end band-select filter provide image rejection? Yes, but since this filter's in-band loss proves critical, its selectivity and hence out-of-band attenuation are inadequate.[6] Thus, a filter with high image rejection typically appears between the LNA and the mixer so that the gain of the LNA lowers the filter's contribution to the receiver noise figure.

The linearity and selectivity required of the image-reject filter have dictated passive, off-chip implementations. Operating at high frequencies, the filters are designed to provide 50-Ω input and output impedances. The LNA must therefore drive a load impedance of 50 Ω, a difficult and power-hungry task.

Image Rejection versus Channel Selection As noted in Fig. 4.13, the desired channel and the image have a frequency difference of $2\omega_{IF}$. Thus, to maximize image rejection, it is desirable to choose a large value for ω_{IF}, i.e., a large difference between ω_{in} and ω_{LO}. How large can $2\omega_{IF}$ be? Recall that the premise in a heterodyne architecture is to translate the center frequency to a sufficiently *low* value so that channel selection by means

6. As mentioned earlier, passive filters suffer from a trade-off between the in-band loss and the out-of-band attenuation.

of practical filters becomes feasible. However, as $2\omega_{IF}$ increases, so does the center of the downconverted channel (ω_{IF}), necessitating a higher Q in the IF filter.

Shown in Fig. 4.14 are two cases corresponding to high and low values of IF so as to illustrate the trade-off. A high IF [Fig. 4.14(a)] allows substantial rejection of the image whereas a low IF [Fig. 4.14(b)] helps with the suppression of in-band interferers. We thus say heterodyne receivers suffer from a trade-off between image rejection and channel selection.

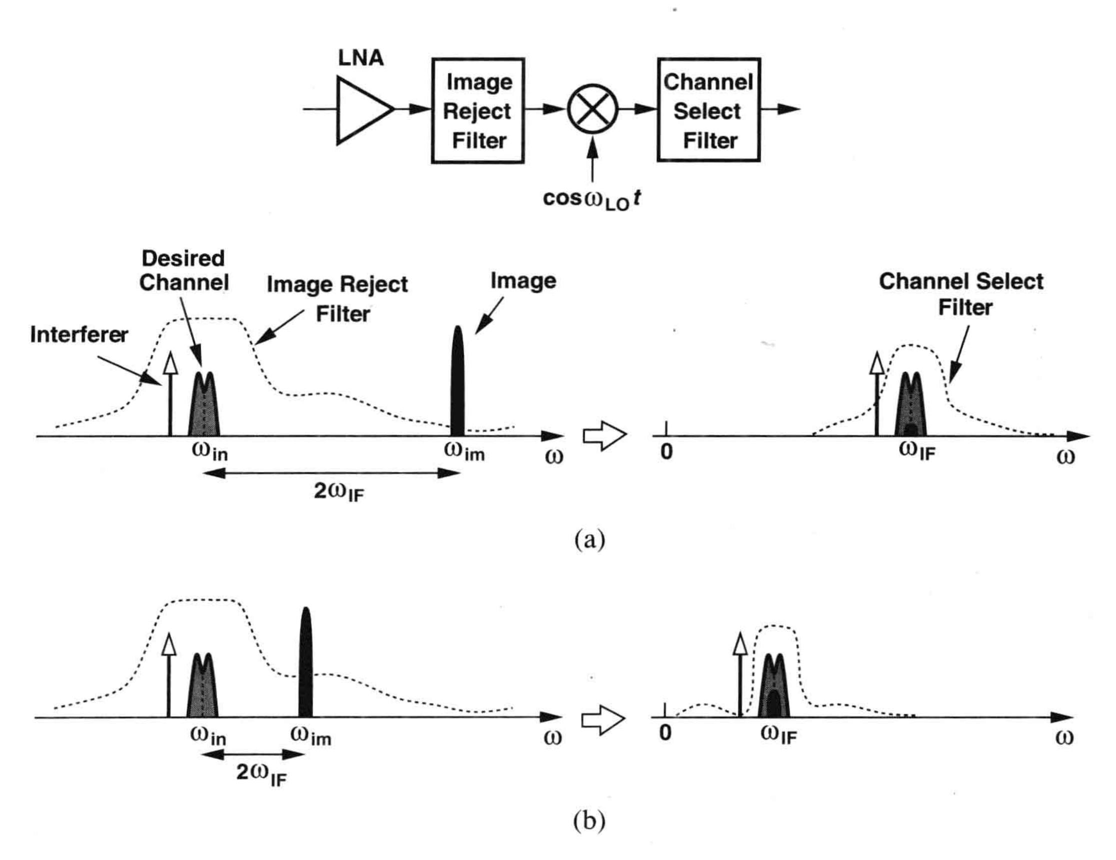


Figure 4.14 *Trade-off between image rejection and channel selection for (a) high IF and (b) low IF.*

Example 4.7

An engineer is to design a receiver for space applications with no concern for interferers. The engineer constructs the heterodyne front end shown in Fig. 4.15(a), avoiding band-select and image-select filters. Explain why this design suffers from a relatively high noise figure.

(Continues)

Example 4.7 (*Continued*)

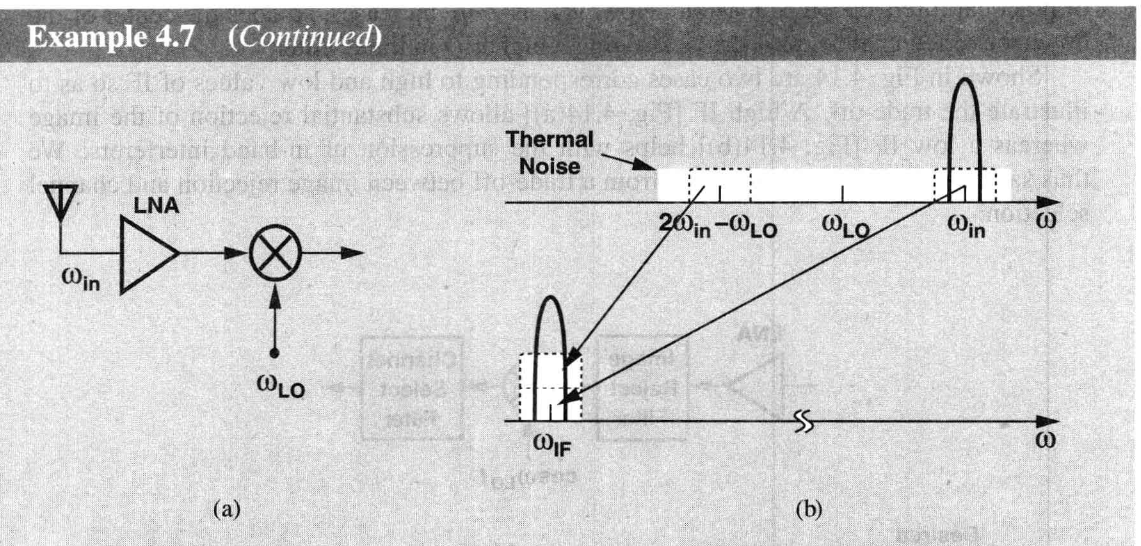


Figure 4.15 *(a) Receiver for space applications, (b) effect of noise in image band.*

Solution:

Even in the absence of interferers, the thermal noise produced by the antenna and the LNA in the image band arrives at the input of the mixer. Thus, the desired signal, the thermal noise in the desired channel, and the thermal noise in the image band are downconverted to IF [Fig. 4.15(b)], leading to a higher noise figure for the receiver (unless the LNA has such a limited bandwidth that it suppresses the noise in the image band). An image-reject filter would remove the noise in the image band. We return to this effect in Chapter 6.

Dual Downconversion The trade-off between image rejection and channel selection in the simple heterodyne architecture of Fig. 4.14 often proves quite severe: if the IF is high, the image can be suppressed but complete channel selection is difficult, and vice versa. To resolve this issue, the concept of heterodyning can be extended to multiple downconversions, each followed by filtering and amplification. Illustrated in Fig. 4.16, this technique performs *partial* channel selection at progressively lower center frequencies, thereby relaxing the Q required of each filter. Note that the second downconversion may also entail an image called the "secondary image" here.

Figure 4.16 also shows the spectra at different points along the cascade. The front-end filter selects the band while providing some image rejection as well. After amplification and image-reject filtering, the spectrum at point C is obtained. A sufficiently linear mixer then translates the desired channel and the adjacent interferers to the first IF (point D). Partial channel selection in BPF_3 permits the use of a second mixer with reasonable linearity. Next, the spectrum is translated to the second IF, and BPF_4 suppresses the interferers to acceptably low levels (point G). We call MX_1 and MX_2 the "RF mixer" and the "IF mixer," respectively.

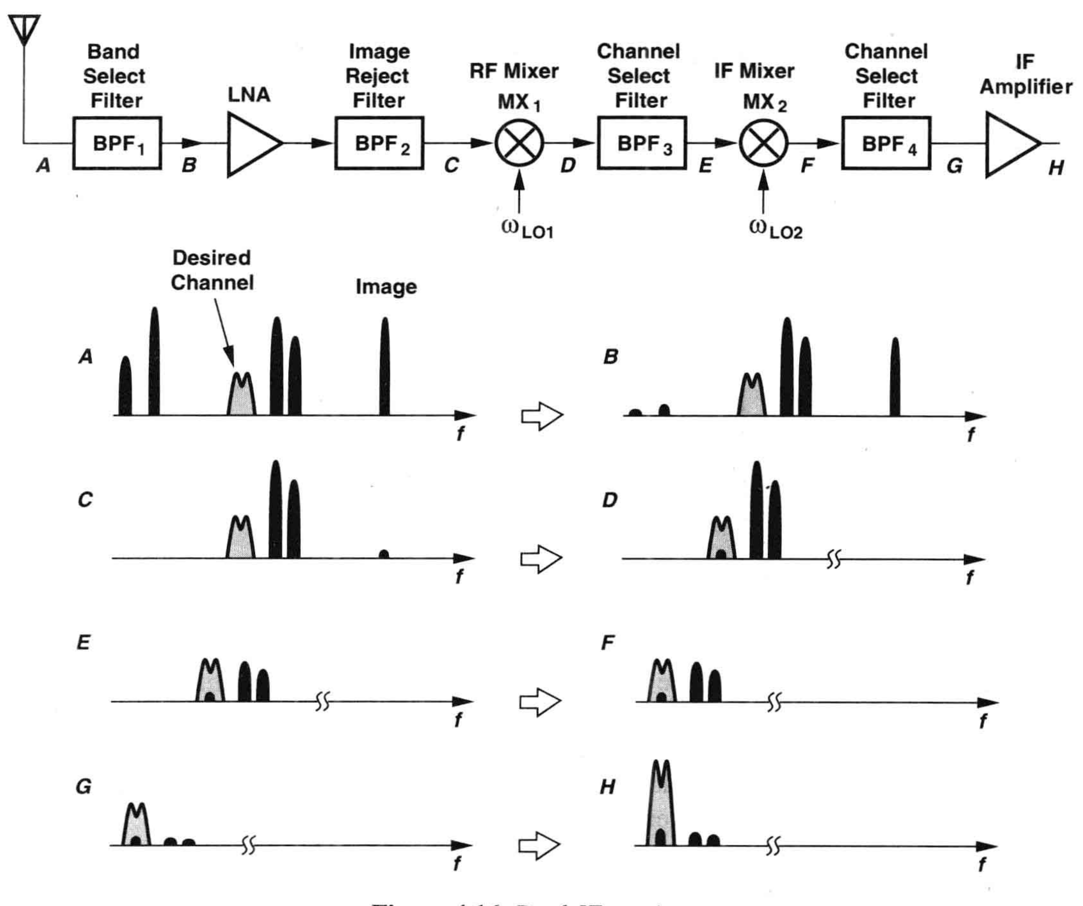


Figure 4.16 *Dual-IF receiver.*

Recall from Chapter 2 that in a cascade of gain stages, the noise figure is most critical in the front end and the linearity in the back end. Thus, an optimum design scales both the noise figure and the IP_3 of each stage according to the total gain preceding that stage. Now suppose the receiver of Fig. 4.16 exhibits a total gain of, say, 40 dB from A to G. If the two IF filters provided *no* channel selection, then the IP_3 of the IF amplifier would need to be about 40 dB higher than that of the LNA, e.g., in the vicinity of +30 dBm. It is difficult to achieve such high linearity with reasonable noise, power dissipation, and gain, especially if the circuit must operate from a low supply voltage. If each IF filter attenuates the in-band interferers to some extent, then the linearity required of the subsequent stages is relaxed proportionally. This is sometimes loosely stated as "every dB of gain requires 1 dB of prefiltering," or "every dB of prefiltering relaxes the IP_3 by 1 dB."

Example 4.8

Assuming low-side injection for both downconversion mixers in Fig. 4.16, determine the image frequencies.

(*Continues*)

Example 4.8 (*Continued*)

Solution:

As shown in Fig. 4.17, the first image lies at $2\omega_{LO1} - \omega_{in}$. The second image is located at $2\omega_{LO2} - (\omega_{in} - \omega_{LO1})$.

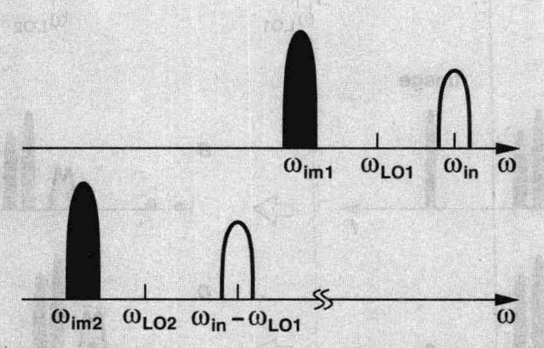


Figure 4.17 *Secondary image in a heterodyne RX.*

Mixing Spurs In the heterodyne receiver of Fig. 4.16, we have assumed ideal RF and IF mixers. In practice, mixers depart from simple analog multipliers, introducing undesirable effects in the receive path. Specifically, as exemplified by the switching mixer studied in Chapter 2, mixers in fact multiply the RF input by a *square-wave* LO even if the LO signal applied to the mixer is a sinusoid. As explained in Chapter 6, this internal sinusoid/square-wave conversion[7] is inevitable in mixer design. We must therefore view mixing as multiplication of the RF input by all harmonics of the LO.[8] In other words, the RF mixer in Fig. 4.16 produces components at $\omega_{in} \pm m\omega_{LO1}$ and the IF mixer, $\omega_{in} \pm m\omega_{LO1} \pm n\omega_{LO2}$, where m and n are integers. For the desired signal, of course, only $\omega_{in} - \omega_{LO1} - \omega_{LO2}$ is of interest. But if an interferer, ω_{int}, is downconverted to the same IF, it corrupts the signal; this occurs if

$$\omega_{int} \pm m\omega_{LO1} \pm n\omega_{LO2} = \omega_{in} - \omega_{LO1} - \omega_{LO2}. \tag{4.7}$$

Called "mixing spurs," such interferers require careful attention to the choice of the LO frequencies.

Example 4.9

Figure 4.18(a) shows a 2.4-GHz dual downconversion receiver, where the first LO frequency is chosen so as to place the (primary) image in the GPS band for some of the channels. Determine a few mixing spurs.

7. Also called "limiting."

8. Or only the odd harmonics of the LO if the LO and the mixer are perfectly symmetric (Chapter 6).

Example 4.9 *(Continued)*

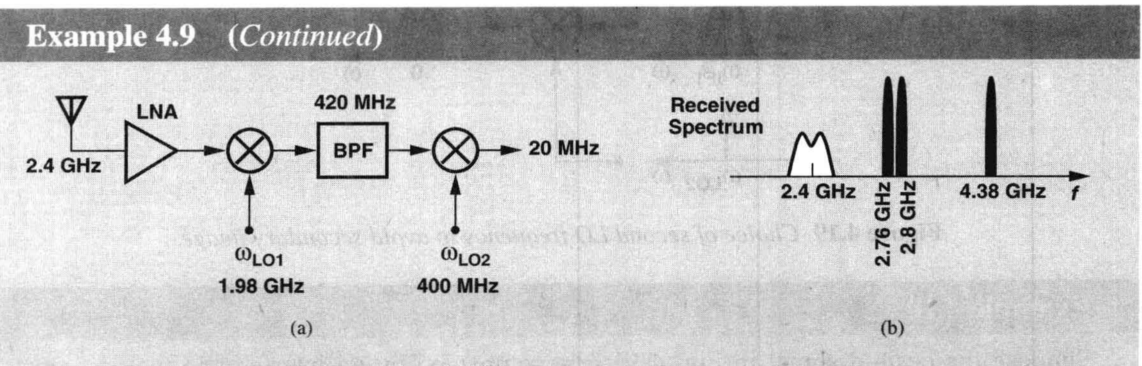


Figure 4.18 *(a) Heterodyne RX for 2.4-GHz band, (b) mixing spurs.*

Solution:

Let us consider the second harmonic of LO_2, 800 MHz. If an interferer appears at the first IF at 820 MHz or 780 MHz, then it coincides with the desired signal at the second IF. In the RF band, the former corresponds to 820 MHz + 1980 MHz = 2.8 GHz and the latter arises from 780 MHz + 1980 MHz = 2.76 GHz. We can also identify the image corresponding to the second harmonic of LO_1 by writing $f_{in} - 2f_{LO1} - f_{LO2} = 20$ MHz and hence $f_{in} = 4.38$ GHz. Figure 4.18(b) summarizes these results. We observe that numerous spurs can be identified by considering other combinations of LO harmonics.

The architecture of Fig. 4.16 consists of two downconversion steps. Is it possible to use more? Yes, but the additional IF filters and LO further complicate the design and, more importantly, the mixing spurs arising from additional downconversion mixers become difficult to manage. For these reasons, most heterodyne receivers employ only two downconversion steps.

4.2.2 Modern Heterodyne Receivers

The receiver of Fig. 4.16 employs several bulky, passive (off-chip) filters and two local oscillators; it has thus become obsolete. Today's architecture and circuit design omits all of the off-chip filters except for the front-end band-select device.

With the omission of a highly-selective filter at the first IF, no channel selection occurs at this point, thereby dictating a high linearity for the second mixer. Fortunately, CMOS mixers achieve high linearities. But the lack of a selective filter also means that the secondary image—that associated with ω_{LO2}—may become serious.

Zero Second IF To avoid the secondary image, most modern heterodyne receivers employ a *zero* second IF. Illustrated in Fig. 4.19, the idea is to place ω_{LO2} at the center of the first IF signal so that the output of the second mixer contains the desired channel with a zero center frequency. In this case, the image is the signal itself, i.e., the left part of the signal spectrum is the image of the right part and vice versa. As explained below, this effect can be handled properly. The key point here is that no interferer at other frequencies can be downconverted as an image to a zero center frequency if $\omega_{LO2} = \omega_{IF1}$.

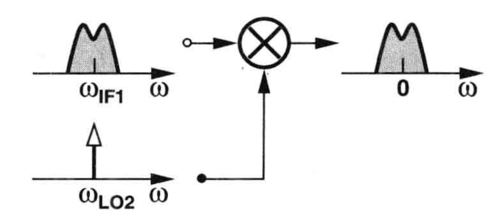


Figure 4.19 *Choice of second LO frequency to avoid secondary image.*

Example 4.10

Suppose the desired signal in Fig. 4.19 is accompanied by an interferer in the adjacent channel. Plot the spectrum at the second IF if $\omega_{LO2} = \omega_{IF1}$.

Solution:

Let us consider the spectra at the first IF carefully. As shown in Fig. 4.20, the desired channel appears at $\pm\omega_{IF1}$ and is accompanied by the interferer.[9] Upon mixing in the time domain, the spectrum at negative frequencies is convolved with the LO impulse at $+\omega_{LO2}$, sliding to a zero center frequency for the desired channel. Similarly, the spectrum at positive frequencies is convolved with the impulse at $-\omega_{LO2}$ and shifted down to zero. The output thus consists of two copies of the desired channel surrounded by the interferer spectrum at both positive and negative frequencies.

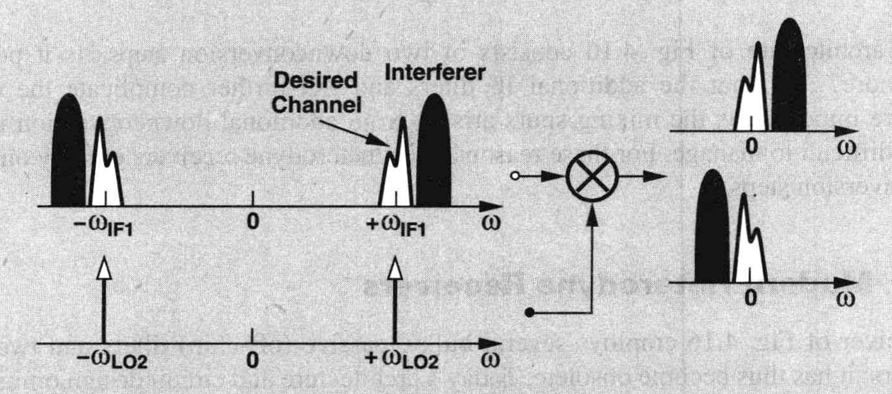


Figure 4.20 *Downconversion of a desired signal and an interferer in the adjacent channel.*

What happens if the signal becomes its own image? To understand this effect, we must distinguish between "symmetrically-modulated" and "asymmetrically-modulated" signals. First, consider the generation of an AM signal, Fig. 4.21(a), where a real baseband signal having a symmetric spectrum $S_a(f)$ is mixed with a carrier, thereby producing an output spectrum that remains symmetric with respect to f_{LO}. We say AM signals are symmetric because their *modulated* spectra carry exactly the same information on both sides of the carrier.[10]

9. The spectrum of a real signal is symmetric with respect to the origin.

10. In fact, it is possible to remove one side without losing information.

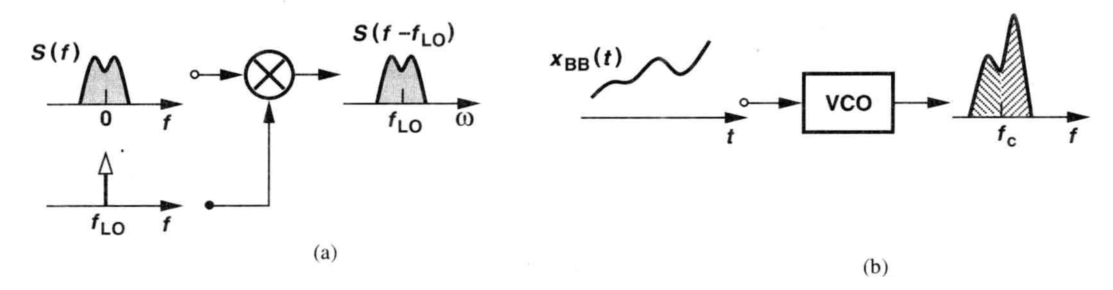


Figure 4.21 *(a) AM signal generation, (b) FM signal generation.*

Now, consider an FM signal generated by a voltage-controlled oscillator [Fig. 4.21(b)] (Chapter 8). We note that as the baseband voltage becomes more positive, the output frequency, say, increases, and vice versa. That is, the information in the sideband below the carrier is different from that in the sideband above the carrier. We say FM signals have an asymmetric spectrum. Most of today's modulation schemes, e.g., FSK, QPSK, GMSK, and QAM, exhibit asymmetric spectra around their carrier frequencies. While the conceptual diagram in Fig. 4.21(b) shows the asymmetry in the *magnitude*, some modulation schemes may exhibit asymmetry in only their phase.

As exemplified by the spectra in Fig. 4.20, downconversion to a zero IF superimposes two copies of the signal, thereby causing corruption if the signal spectrum is asymmetric. Figure 4.22 depicts this effect more explicitly.

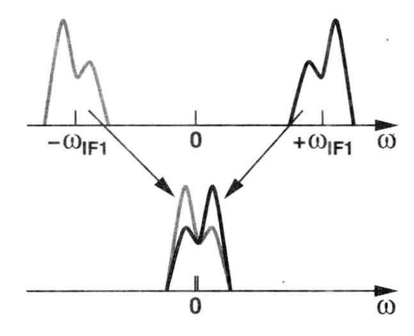


Figure 4.22 *Overlap of signal sidebands after second downconversion.*

Example 4.11

Downconversion to what minimum intermediate frequency avoids self-corruption of asymmetric signals?

Solution:

To avoid self-corruption, the downconverted spectra must not overlap each other. Thus, as shown in Fig. 4.23, the signal can be downconverted to an IF equal to *half* of the signal bandwidth. Of course, an interferer may now become the image.

(Continues)

Example 4.11 (*Continued*)

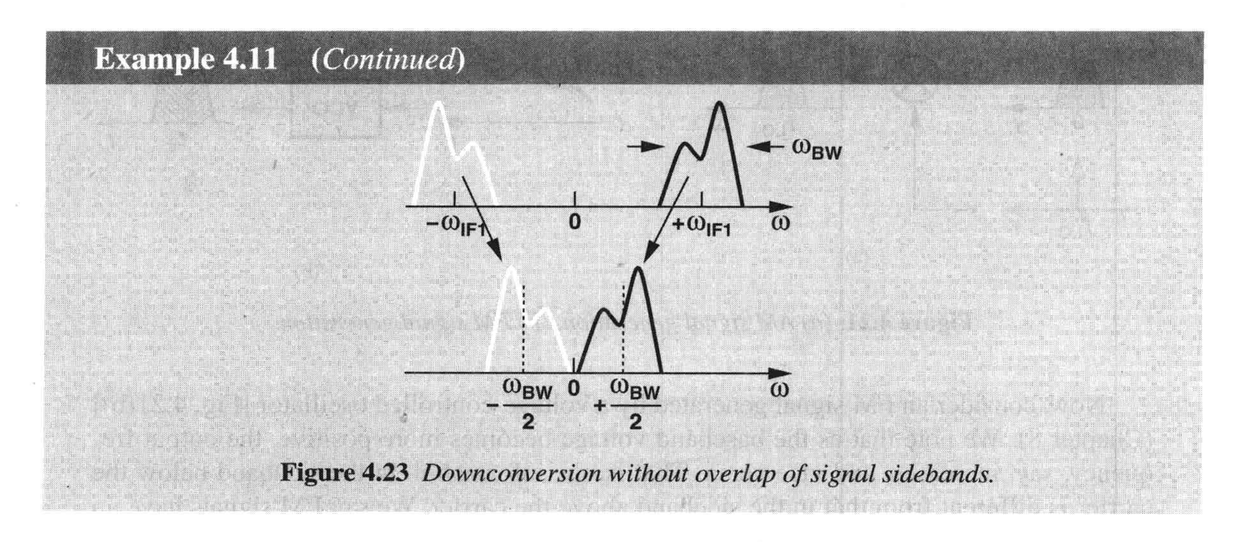


Figure 4.23 *Downconversion without overlap of signal sidebands.*

How can downconversion to a zero IF avoid self-corruption? This is accomplished by creating *two* versions of the downconverted signal that have a phase difference of 90°. Illustrated in Fig. 4.24, "quadrature downconversion" is performed by mixing $x_{IF}(t)$ with the quadrature phases of the second LO ($\omega_{LO2} = \omega_{IF1}$). The resulting outputs, $x_{BB,I}(t)$ and $x_{BB,Q}(t)$, are called the "quadrature baseband signals." Though exhibiting identical spectra, $x_{BB,I}(t)$ and $x_{BB,Q}(t)$ are separated in phase and together can reconstruct the original information. In Problem 4.8, we show that even an AM signal of the form $A(t)\cos\omega_c t$ may require quadrature downconversion.

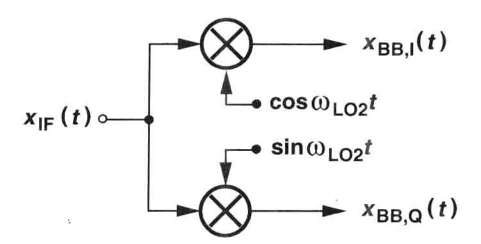


Figure 4.24 *Quadrature downconversion.*

Figure 4.25 shows a heterodyne receiver constructed after the above principles. In the absence of an (external) image-reject filter, the LNA need not drive a 50-Ω load, and the LNA/mixer interface can be optimized for gain, noise, and linearity with little concern for the interface impedance values. However, the lack of an image-reject filter requires careful attention to the interferers in the image band, and dictates a narrow-band LNA design so that the thermal noise of the antenna and the LNA in the image band is heavily suppressed (Example 4.7). Moreover, no channel-select filter is shown at the first IF, but some "mild" on-chip band-pass filtering is usually inserted here to suppress out-of-band interferers. For example, the RF mixer may incorporate an LC load, providing some filtration.

Sliding-IF Receivers Modern heterodyne receivers entail another important departure from their older counterparts: they employ only one oscillator. This is because the design of oscillators and frequency synthesizers proves difficult and, more importantly, oscillators

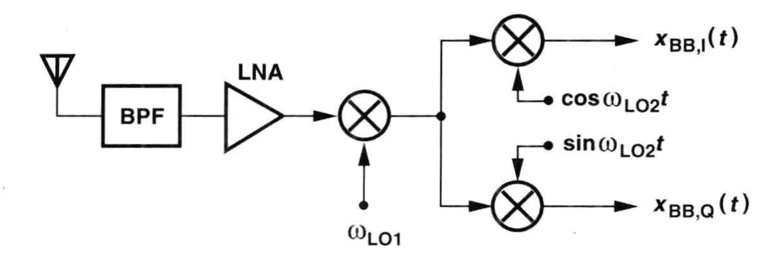


Figure 4.25 *Heterodyne RX with quadrature downconversion.*

Figure 4.26 *(a) Sliding-IF heterodyne RX, (b) divide-by-2 circuit waveforms, (c) resulting spectra.*

fabricated on the same chip suffer from unwanted coupling. The second LO frequency is therefore *derived* from the first by "frequency division."[11] Shown in Fig. 4.26(a) is an example, where the first LO is followed by a $\div 2$ circuit to generate the second LO waveforms at a frequency of $f_{LO1}/2$. As depicted in Fig. 4.26(b) and explained in Chapter 10, certain $\div 2$ topologies can produce quadrature outputs. Figure 4.26(c) shows the spectra at different points in the receiver.

11. Frequency division can be performed by a counter: for M input cycles, the counter produces one output cycle.

The receiver architecture of Fig. 4.26(a) has a number of interesting properties. To translate an input frequency of f_{in} to a second IF of zero, we must have

$$f_{LO1} + \frac{1}{2}f_{LO1} = f_{in} \tag{4.8}$$

and hence

$$f_{LO1} = \frac{2}{3}f_{in}. \tag{4.9}$$

That is, for an input band spanning the range $[f_1\ f_2]$, the LO must cover a range of $[(2/3)f_1\ (2/3)f_2]$ (Fig. 4.27). Moreover, the first IF in this architecture is *not* constant because

$$f_{IF1} = f_{in} - f_{LO} \tag{4.10}$$

$$= \frac{1}{3}f_{in}. \tag{4.11}$$

Thus, as f_{in} varies from f_1 to f_2, f_{IF1} goes from $f_1/3$ to $f_2/3$ (Fig. 4.27). For this reason, this topology is called the "sliding-IF architecture." Unlike the conventional heterodyne receiver of Fig. 4.16, where the first IF filter must exhibit a *narrow* bandwidth to perform some channel selection, this sliding IF topology requires a fractional (or normalized) IF bandwidth[12] equal to the RF input fractional bandwidth. This is because the former is given by

$$\mathrm{BW}_{IF,frac} = \frac{\frac{1}{3}f_2 - \frac{1}{3}f_1}{\left(\frac{1}{3}f_2 + \frac{1}{3}f_1\right)/2}, \tag{4.12}$$

and the latter,

$$\mathrm{BW}_{RF,frac} = \frac{f_2 - f_1}{(f_2 + f_1)/2}. \tag{4.13}$$

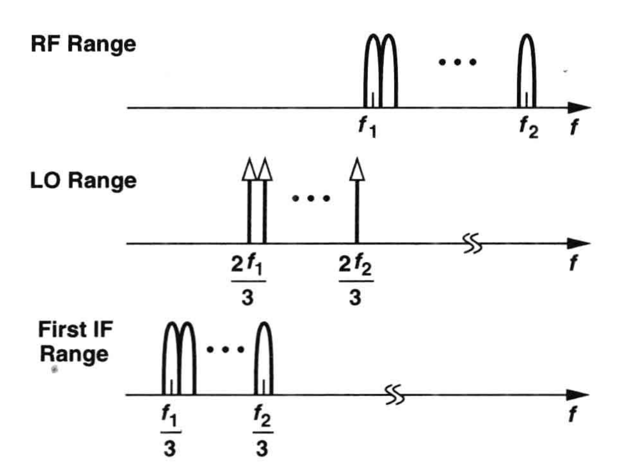


Figure 4.27 *LO and IF ranges in the sliding-IF RX.*

12. Fractional bandwidth is defined as the bandwidth of interest divided by the center frequency of the band.

Example 4.12

Suppose the input band is partitioned into N channels, each having a bandwidth of $(f_2 - f_1)/N = \Delta f$. How does the LO frequency vary as the receiver translates each channel to a zero second IF?

Solution:

The first channel is located between f_1 and $f_1 + \Delta f$. Thus the first LO frequency is chosen equal to two-thirds of the *center* of the channel: $f_{LO} = (2/3)(f_1 + \Delta f/2)$. Similarly, for the second channel, located between $f_1 + \Delta f$ and $f_1 + 2\Delta f$, the LO frequency must be equal to $(2/3)(f_1 + 3\Delta f/2)$. In other words, the LO increments in steps of $(2/3)\Delta f$.

Example 4.13

With the aid of the frequency bands shown in Fig. 4.27, determine the image band for the architecture of Fig. 4.26(a).

Solution:

For an LO frequency of $(2/3)f_1$, the image lies at $2f_{LO} - f_{in} = f_1/3$. Similarly, if $f_{LO1} = (2/3)f_2$, then the image is located at $f_2/3$. Thus, the image band spans the range $[f_1/3\ f_2/3]$ (Fig. 4.28). Interestingly, the image band is *narrower* than the input band.

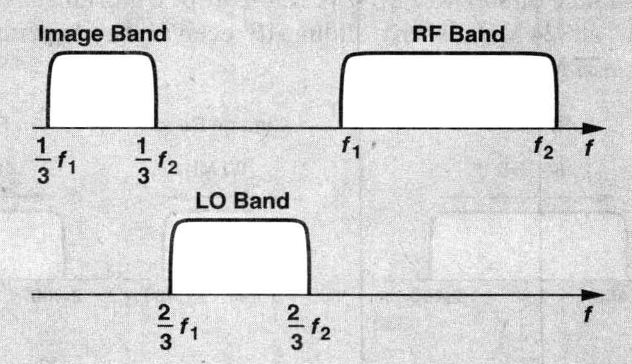


Figure 4.28 *Image band in the sliding-IF RX.*

Does this mean that the image for each channel is also narrower? No, recall from the above example that the LO increments by $(2/3)\Delta f$ as we go from one channel to the next. Thus, consecutive image channels have an overlap of $\Delta f/3$.

The sliding-IF architecture may incorporate greater divide ratios in the generation of the second LO from the first. For example, a $\div 4$ circuit produces quadrature outputs at $f_{LO1}/4$, leading to the following relationship

$$f_{LO1} + \frac{1}{4} f_{LO1} = f_{in} \tag{4.14}$$

and hence

$$f_{LO1} = \frac{4}{5} f_{in}. \tag{4.15}$$

The detailed spectra of such an architecture are studied in Problem 4.1. But we must make two observations here. (1) With a ÷4 circuit, the second LO frequency is equal to $f_{in}/5$, slightly lower than that of the first sliding-IF architecture. This is desirable because generation of LO quadrature phases at lower frequencies incurs smaller mismatches. (2) Unfortunately, the use of a ÷4 circuit reduces the frequency difference between the image and the signal, making it more difficult to reject the image and even the thermal noise of the antenna and the LNA in the image band. In other words, the choice of the divide ratio is governed by the trade-off between quadrature accuracy and image rejection.

Example 4.14

We wish to select a sliding-IF architecture for an 802.11g receiver. Determine the pros and cons of a ÷2 or a ÷4 circuit in the LO path.

Solution:

With a ÷2 circuit, the 11g band (2.40–2.48 GHz) requires an LO range of 1.600–1.653 GHz and hence an image range of 800–827 MHz [Fig. 4.29(a)]. Unfortunately, since the CDMA transmit band begins at 824 MHz, such a sliding-IF receiver may experience a large image in the range of 824–827 MHz.

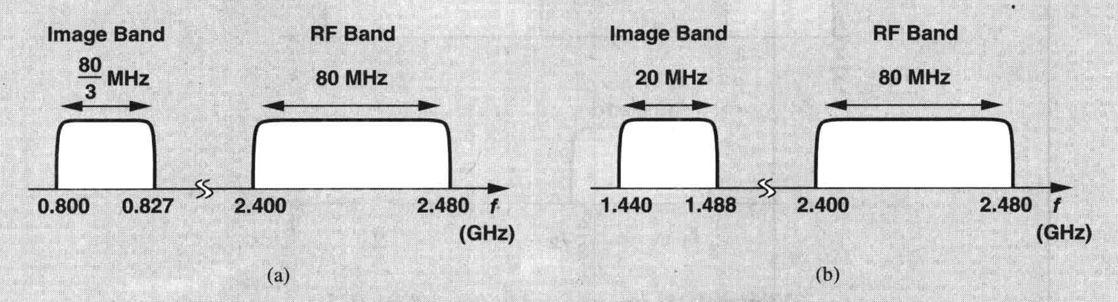


Figure 4.29 *Image band in an 11g RX with a (a) divide-by-2 circuit, (b) divide-by-4 circuit.*

With a ÷4 circuit, the LO range is 1.920–1.984 GHz and the image range, 1.440–1.488 GHz [Fig. 4.29(b)]. This image band is relatively quiet. (Only Japan has allocated a band around 1.4 GHz to WCDMA.) Thus, the choice of the ÷4 ratio proves advantageous here if the LNA selectivity can suppress the thermal noise in the image band. The first IF is lower in the second case and may be beneficial in some implementations.

The baseband signals produced by the heterodyne architecture of Fig. 4.26(a) suffer from a number of critical imperfections, we study these effects in the context of direct-conversion architectures.

4.2.3 Direct-Conversion Receivers

In our study of heterodyne receivers, the reader may have wondered why the RF spectrum is not simply translated to the baseband in the first downconversion. Called the "direct-conversion," "zero-IF," or "homodyne" architecture,[13] this type of receiver entails its own issues but has become popular in the past decade. As explained in Section 4.2.2 and illustrated in Fig. 4.22, downconversion of an asymmetrically-modulated signal to a zero IF leads to self-corruption unless the baseband signals are separated by their phases. The direct-conversion receiver (DCR) therefore emerges as shown in Fig. 4.30, where $\omega_{LO} = \omega_{in}$.

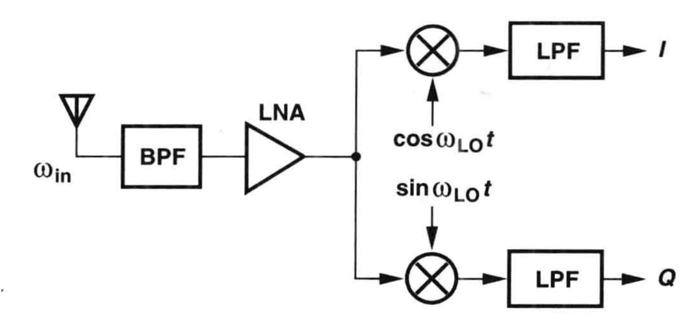


Figure 4.30 *Direct-conversion receiver.*

Three aspects of direct conversion make it a superior choice with respect to heterodyning. First, the absence of an image greatly simplifies the design process. Second, channel selection is performed by *low-pass* filters, which can be realized on-chip as active circuit topologies with relatively sharp cut-off characteristics. Third, mixing spurs are considerably reduced in number and hence simpler to handle.

The architecture of Fig. 4.30 appears to easily lend itself to integration. Except for the front-end band-select filter, the cascade of stages need not connect to external components, and the LNA/mixer interface can be optimized for gain, noise, and linearity without requiring a 50-Ω impedance. The simplicity of the architecture motivated many attempts in the history of RF design, but it was only in the 1990s and 2000s that integration and sophisticated signal processing made direct conversion a viable choice. We now describe the issues that DCRs face and introduce methods of resolving them. Many of these issues also appear in *heterodyne* receivers having a zero second IF.

LO Leakage A direct-conversion receiver *emits* a fraction of its LO power from its antenna. To understand this effect, consider the simplified topology shown in Fig. 4.31, where the LO couples to the antenna through two paths: (1) device capacitances between the LO and RF ports of the mixer and device capacitances or resistances between the output and input of the LNA; (2) the substrate to the input pad, especially because the LO employs large on-chip spiral inductors. The LO emission is undesirable because it may desensitize other receivers operating in the same band. Typical acceptable values range from −50 to −70 dBm (measured at the antenna).

13. The term *homodyne* originates from *homo* (same) and *dyne* (mixing) and has been historically used for only "coherent" reception.

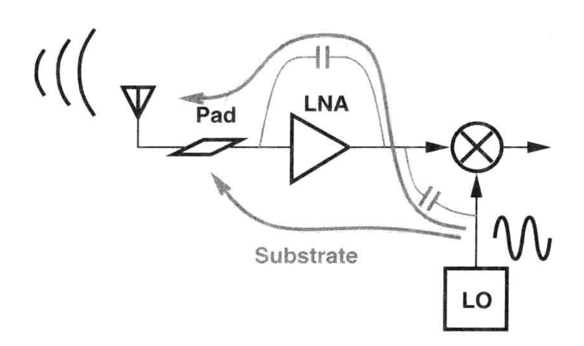


Figure 4.31 *LO leakage.*

Example 4.15

Determine the LO leakage from the output to the input of a cascode LNA.

Solution:

As depicted in Fig. 4.32(a), we apply a test voltage to the output and measure the voltage delivered to the antenna, R_{ant}. Considering only r_{O2} and C_{GD1} as the leakage path, we

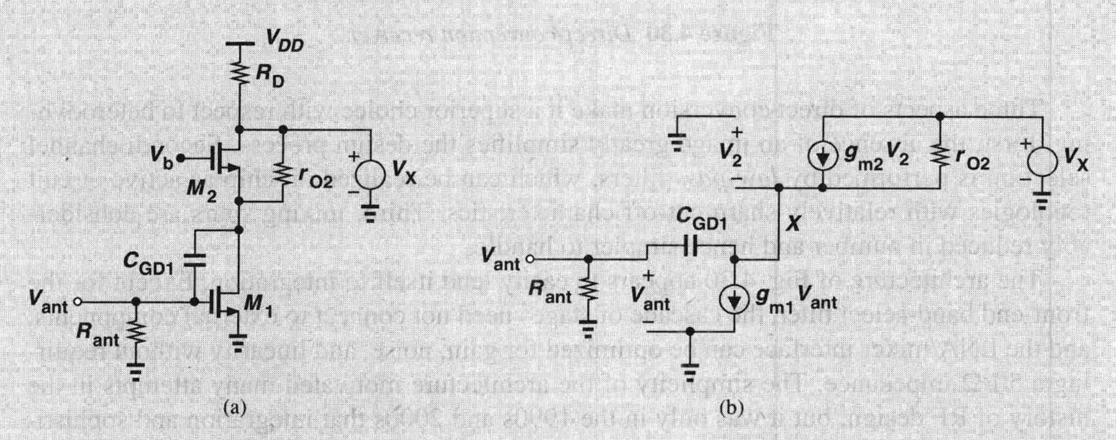


Figure 4.32 *LO leakage in a cascode LNA.*

construct the equivalent circuit shown in Fig. 4.32(b), note that the current flowing through R_{ant} and C_{GD1} is given by V_{ant}/R_{ant}, and write $V_2 = -[V_{ant} + V_{ant}/(R_{ant}C_{GD1}s)]$. Thus, a KCL at node X yields

$$\left(V_{ant} + \frac{V_{ant}}{R_{ant}C_{GD1}s}\right)g_{m2} + \frac{V_{ant}}{R_{ant}} + g_{m1}V_{ant} = \frac{1}{r_{O2}}\left[V_X - \left(V_{ant} + \frac{V_{ant}}{R_{ant}C_{GD1}s}\right)\right]. \quad (4.16)$$

If $g_{m2} \gg 1/r_{O2}$,

$$\frac{V_{ant}}{V_X} \approx \frac{C_{GD1}s}{(g_{m1}R_{ant} + g_{m2}R_{ant} + 1)C_{GD1}s + g_{m2}} \cdot \frac{R_{ant}}{r_{O2}}. \quad (4.17)$$

Example 4.15 (*Continued*)

This quantity is called the "reverse isolation" of the LNA. In a typical design, the denominator is approximately equal to g_{m2}, yielding a value of $R_{ant}C_{GD1}\omega/(g_{m2}r_{O2})$ for V_{out}/V_X.

Does LO leakage occur in heterodyne receivers? Yes, but since the LO frequency falls *outside* the band, it is suppressed by the front-end band-select filters in both the emitting receiver and the victim receiver.

LO leakage can be minimized through symmetric layout of the oscillator and the RF signal path. For example, as shown in Fig. 4.33, if the LO produces differential outputs and the leakage paths from the LO to the input pad remain symmetric, then no LO is emitted from the antenna. In other words, LO leakage arises primarily from random or deterministic *asymmetries* in the circuits and the LO waveform.

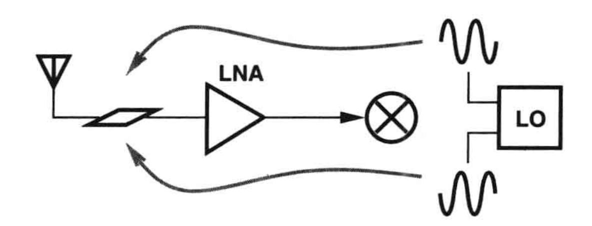


Figure 4.33 *Cancellation of LO leakage by symmetry.*

DC Offsets The LO leakage phenomenon studied above also gives rise to relatively large dc offsets in the baseband, thus creating certain difficulties in the design. Let us first see how the dc offset is generated. Consider the simplified receiver in Fig. 4.34, where a finite amount of in-band LO leakage, kV_{LO}, appears at the LNA input. Along with the desired signal, V_{RF}, this component is amplified and mixed with the LO. Called "LO self-mixing," this effect produces a dc component in the baseband because multiplying a sinusoid by itself results in a dc term.

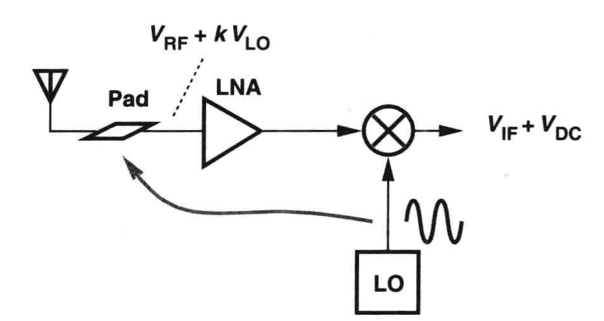


Figure 4.34 *DC offset in a direct-conversion RX.*

Why is a dc component troublesome? It appears that, if constant, a dc term does not corrupt the desired signal. However, such a component makes the processing of the baseband signal difficult. To appreciate the issue, we make three observations: (1) the cascade

of RF and baseband stages in a receiver must amplify the antenna signal by typically 70 to 100 dB; (as a rule of thumb, the signal at the end of the baseband chain should reach roughly 0 dBm.) (2) the received signal and the LO leakage are amplified and processed alongside each other; (3) for an RF signal level of, say, −80 dBm at the antenna, the receiver must provide a gain of about 80 dB, which, applied to an LO leakage of, say, −60 dBm, yields a very large dc offset in the baseband stages. Such an offset saturates the baseband circuits, simply prohibiting signal detection.

Example 4.16

A direct-conversion receiver incorporates a voltage gain of 30 dB from the LNA input to each mixer output and another gain of 40 dB in the baseband stages following the mixer (Fig. 4.35). If the LO leakage at the LNA input is equal to −60 dBm, determine the offset voltage at the output of the mixer and at the output of the baseband chain.

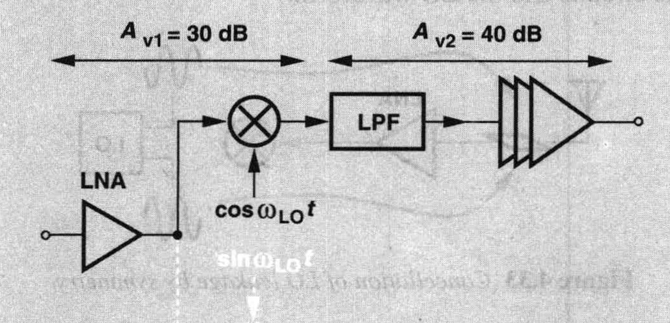


Figure 4.35 *Effect of dc offset in baseband chain.*

Solution:

What does $A_{V1} = 30$ dB mean? If a sinusoid $V_0 \cos\omega_{in}t$ is applied to the LNA input, then the baseband signal at the mixer output, $V_{bb}\cos(\omega_{in} - \omega_{LO})t$, has an amplitude given by

$$V_{bb} = A_{V1} \cdot V_0. \tag{4.18}$$

Thus, for an input $V_{leak}\cos\omega_{LO}t$, the dc value at the mixer output is equal to

$$V_{dc} = A_{V1} \cdot V_{leak}. \tag{4.19}$$

Since $A_{V1} = 31.6$ and $V_{leak} = (632/2)\ \mu$V, we have $V_{dc} = 10$ mV. Amplified by another 40 dB, this offset reaches 1 V at the baseband output!

Example 4.17

The dc offsets measured in the baseband I and Q outputs are often *unequal*. Explain why.

Example 4.17 (*Continued*)

Solution:

Suppose, in the presence of the quadrature phases of the LO, the net LO leakage at the input of the LNA is expressed as $V_{leak} \cos(\omega_{LO} t + \phi_{leak})$, where ϕ_{leak} arises from the phase shift through the path(s) from each LO phase to the LNA input and also the summation of the leakages $V_{LO} \cos \omega_{LO} t$ and $V_{LO} \sin \omega_{LO} t$ (Fig. 4.36). The LO leakage travels through the LNA and each mixer, experiencing an additional phase shift, ϕ_{ckt}, and is multiplied by $V_{LO} \cos \omega_{LO} t$ and $V_{LO} \sin \omega_{LO} t$. The dc components are therefore given by

$$V_{dc,I} = \alpha V_{leak} V_{LO} \cos(\phi_{leak} + \phi_{ckt}) \tag{4.20}$$

$$V_{dc,Q} = -\alpha V_{leak} V_{LO} \sin(\phi_{leak} + \phi_{ckt}). \tag{4.21}$$

Thus, the two dc offsets are generally unequal.

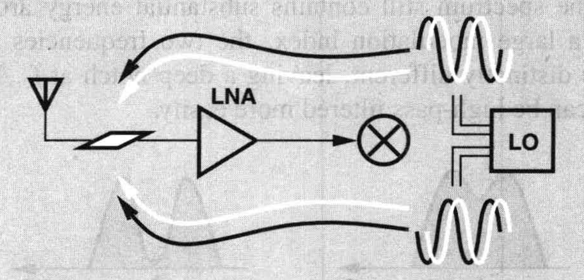


Figure 4.36 *Leakage of quadrature phases of LO.*

Does the problem of dc offsets occur in heterodyne receivers having a zero second IF [Fig. 4.26(a)]? Yes, the leakage of the second LO to the input of the IF mixers produces dc offsets in the baseband. Since the second LO frequency is equal to $f_{in}/3$ in Fig. 4.26(a), the leakage is smaller than that in direct-conversion receivers,[14] but the dc offset is still large enough to saturate the baseband stages or at least create substantial nonlinearity.

The foregoing study implies that receivers having a final zero IF must incorporate some means of offset cancellation in each of the baseband I and Q paths. A natural candidate is a high-pass filter (ac coupling) as shown in Fig. 4.37(a), where C_1 blocks the dc offset and R_1 establishes proper bias, V_b, for the input of A_1. However, as depicted in Fig. 4.37(b), such a network also removes a fraction of the signal's spectrum near zero frequency, thereby introducing intersymbol interference. As a rule of thumb, the corner frequency of the high-pass filter, $f_1 = (2\pi R_1 C_1)^{-1}$, must be less than one-thousandth of the symbol rate for negligible ISI. In practice, careful simulations are necessary to determine the maximum tolerable value of f_1 for a given modulation scheme.

The feasibility of on-chip ac coupling depends on both the symbol rate and the type of modulation. For example, the bit rate of 271 kb/s in GSM necessitates a corner frequency of roughly 20–30 Hz and hence extremely large capacitors and/or resistors. Note

14. Also because the LO in direct-conversion receivers employs inductors, which couple the LO waveform into the substrate, whereas the second LO in heterodyne architectures is produced by an inductor-less divider.

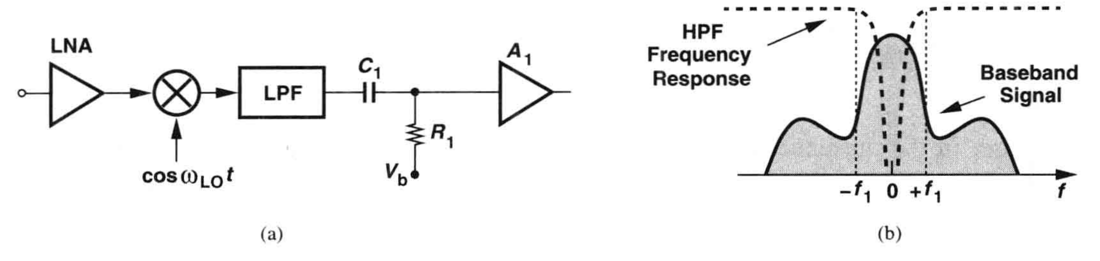


Figure 4.37 *(a) Use of a high-pass filter to remove dc offset, (b) effect on signal spectrum.*

that the quadrature mixers require *four* high-pass networks in their differential outputs. On the other hand, 802.11b at a maximum bit rate of 20 Mb/s can operate with a high-pass corner frequency of 20 kHz, a barely feasible value for on-chip integration.

Modulation schemes that contain little energy around the carrier better lend themselves to ac coupling in the baseband. Figure 4.38 depicts two cases for FSK signals: for a small modulation index, the spectrum still contains substantial energy around the carrier frequency, f_c, but for a large modulation index, the two frequencies generated by ONEs and ZEROs become distinctly different, leaving a deep notch at f_c. If downconverted to baseband, the latter can be high-pass filtered more easily.

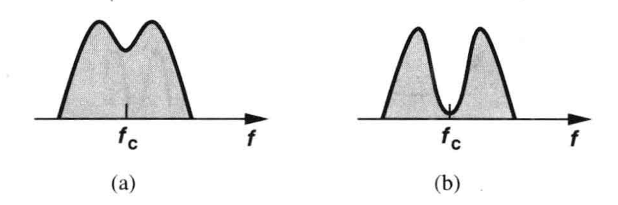


Figure 4.38 *FSK spectrum with (a) small and (b) large frequency deviation.*

A drawback of ac coupling stems from its slow response to transient inputs. With a very low $f_1 = (2\pi R_1 C_1)^{-1}$, the circuit inevitably suffers from a long time constant, failing to block the offset if the offset suddenly changes. This change occurs if (a) the LO frequency is switched to another channel, hence changing the LO leakage, or (b) the *gain* of the LNA is switched to a different value, thus changing the reverse isolation of the LNA. (LNA gain switching is necessary to accommodate varying levels of the received signal.) For these reasons, and due to the relatively large size of the required capacitors, ac coupling is rarely used in today's direct-conversion receivers.

Example 4.18

Figure 4.39(a) shows another method of suppressing dc offsets in the baseband. Here, the main signal path consists of G_{m1} (a transconductance amplifier), R_D, and A_1, providing a total voltage gain of $G_{m1}R_DA_1$. The negative feedback branch comprising R_1, C_1 and $-G_{mF}$ returns a low-frequency current to node X so as to drive the dc content of V_{out} toward zero. Note that this topology suppresses the dc offsets of all of the stages in the baseband. Calculate the corner frequency of the circuit.

Example 4.18 *(Continued)*

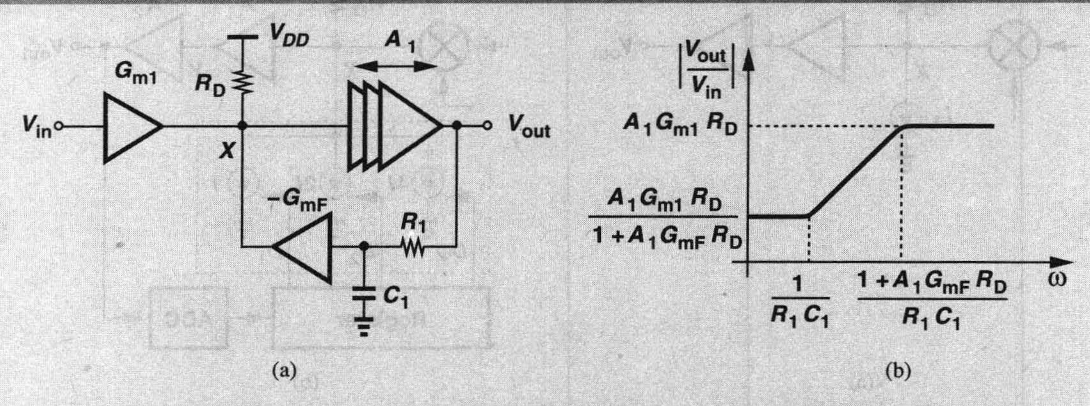


Figure 4.39 *(a) Offset cancellation by feedback, (b) resulting frequency response.*

Solution:

Recognizing that the current returned by $-G_{mF}$ to node X is equal to $-G_{mF}V_{out}/(R_1C_1s + 1)$ and the current produced by G_{m1} is given by $G_{m1}V_{in}$, we sum the two at node X, multiply the sum by R_D and A_1, and equate the result to V_{out}

$$\left(\frac{-G_{mF}V_{out}}{R_1C_1s + 1} + G_{m1}V_{in}\right)R_DA_1 = V_{out}. \tag{4.22}$$

It follows that

$$\frac{V_{out}}{V_{in}} = \frac{G_{m1}R_DA_1(R_1C_1s + 1)}{R_1C_1s + G_{mF}R_DA_1 + 1}. \tag{4.23}$$

The circuit thus exhibits a pole at $-(1 + G_{mF}R_DA_1)/(R_1C_1)$ and a zero at $-1/(R_1C_1)$ [Fig. 4.39(b)]. The input offset is amplified by a factor of $G_{m1}R_DA_1/(1 + G_{mF}R_DA_1) \approx G_{m1}/G_{mF}$ if $G_{mF}R_DA_1 \gg 1$. This gain must remain below unity, i.e., G_{mF} is typically chosen larger than G_{m1}. Unfortunately, the high-pass corner frequency is given by

$$f_1 \approx \frac{G_{mF}R_DA_1}{2\pi(R_1C_1)}, \tag{4.24}$$

a factor of $G_{mF}R_DA_1$ higher than that of the passive circuit in Fig. 4.37(a). This "active feedback" circuit therefore requires greater values for R_1 and C_1 to provide a low f_1. The advantage is that C_1 can be realized by a MOSFET [while that in Fig. 4.37(a) cannot].

The most common approach to offset cancellation employs digital-to-analog converters (DACs) to draw a corrective current in the same manner as the G_{mF} stage in Fig. 4.39(a). Let us first consider the cascade shown in Fig. 4.40(a), where I_1 is drawn from node X

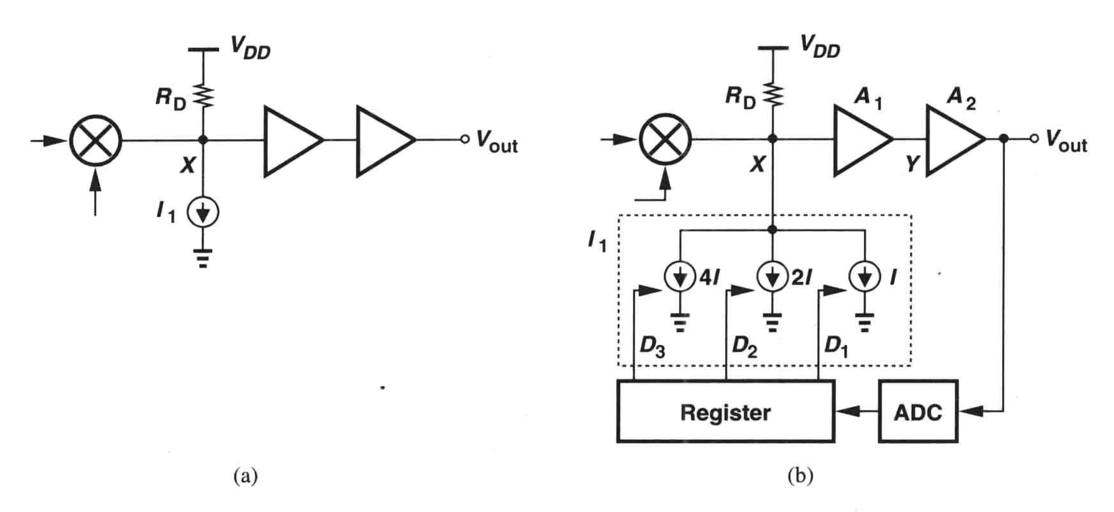


Figure 4.40 *(a) Offset cancellation by means of a current source, (b) actual implementation.*

and its value is adjusted so as to drive the dc content in V_{out} to zero.[15] For example, if the mixer produces an offset of ΔV at X and the subsequent stages exhibit no offset, then $I_1 = \Delta V/R_D$ with proper polarity. In Fig. 4.39(a), the corrective current provided by G_{mF} is continuously adjusted (even in the presence of signal), leading to the high-pass behavior; we thus seek a method of "freezing" the value of I_1 so that it does not affect the baseband frequency response. This requires that I_1 be controlled by a register and hence vary in discrete steps. As illustrated in Fig. 4.40(b), I_1 is decomposed into units that are turned on or off according to the values stored in the register. For example, a binary word $D_3D_2D_1$ controls "binary-weighted" current sources $4I$, $2I$, and I. These current sources form a DAC.

How is the correct value of the register determined? When the receiver is turned on, an analog-to-digital converter (ADC) digitizes the baseband output (in the absence of signals) and drives the register. The entire negative-feedback loop thus converges such that V_{out} is minimized. The resulting values are then stored in the register and remain frozen during the actual operation of the receiver.

The arrangement of Fig. 4.40(b) appears rather complex, but, with the scaling of CMOS technology, the area occupied by the DAC and the register is in fact considerably *smaller* than that of the capacitors in Figs. 4.37(a) and 4.39(a). Moreover, the ADC is also used during signal reception.

The digital storage of offset affords other capabilities as well. For example, since the offset may vary with the LO frequency or gain settings before or after the mixer, at power-up the receiver is cycled through all possible combinations of LO and gain settings, and the required values of I_1 are stored in a small memory. During reception, for the given LO and gain settings, the value of I_1 is recalled from the memory and loaded into the register.

The principal drawback of digital storage originates from the finite resolution with which the offset is cancelled. For example, with the 3-bit DAC in Fig. 4.40(b), an offset of, say, 10 mV at node X, can be reduced to about 1.2 mV after the overall loop settles. Thus, for an A_1A_2 of, say, 40 dB, V_{out} still suffers from 120 mV of offset. To alleviate this issue,

15. We assume that the mixer generates an output *current*.

a higher resolution must be realized or multiple DACs must be tied to different nodes (e.g., Y and V_{out}) in the cascade to limit the maximum offset.

Example 4.19

In the arrangement of Fig. 4.40(b), another 3-bit DAC is tied to node Y. If the mixer produces an offset of 10 mV and $A_1A_2 = 40$ dB, what is the minimum offset that can be achieved in V_{out}? Assume A_1 and A_2 have no offset.

Solution:

The second DAC lowers the output offset by another factor of 8, yielding a minimum of about 10 mV $\times$ 100/64 $\approx$ 16 mV.

Even-Order Distortion Our study of nonlinearity in Chapter 2 indicates that third-order distortion results in compression and intermodulation. Direct-conversion receivers are additionally sensitive to even-order nonlinearity in the RF path, and so are heterodyne architectures having a second zero IF.

Suppose, as shown in Fig. 4.41, two strong interferers at ω_1 and ω_2 experience a nonlinearity such as $y(t) = \alpha_1 x(t) + \alpha_2 x^2(t)$ in the LNA. The second-order term yields the product of these two interferers and hence a low-frequency "beat" at $\omega_2 - \omega_1$. What is the effect of this component? Upon multiplication by $\cos \omega_{LO} t$ in an *ideal* mixer, such a term is translated to high frequencies and hence becomes unimportant. In reality, however, asymmetries in the mixer or in the LO waveform allow a fraction of the RF input of the mixer to appear at the output *without* frequency translation. As a result, a fraction of the low-frequency beat appears in the baseband, thereby corrupting the downconverted signal. Of course, the beat generated by the LNA can be removed by ac coupling, making the input transistor of the *mixer* the dominant source of even-order distortion.

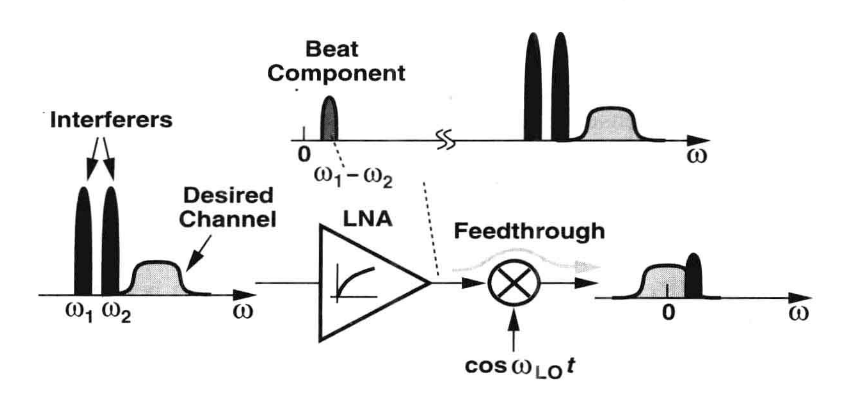


Figure 4.41 *Effect of even-order distortion on direct conversion.*

To understand how asymmetries give rise to direct "feedthrough" in a mixer, first consider the circuit shown in Fig. 4.42(a). As explained in Chapter 2, the output can be written as the product of V_{in} and an ideal LO, i.e., a square-wave toggling between 0 and 1 with

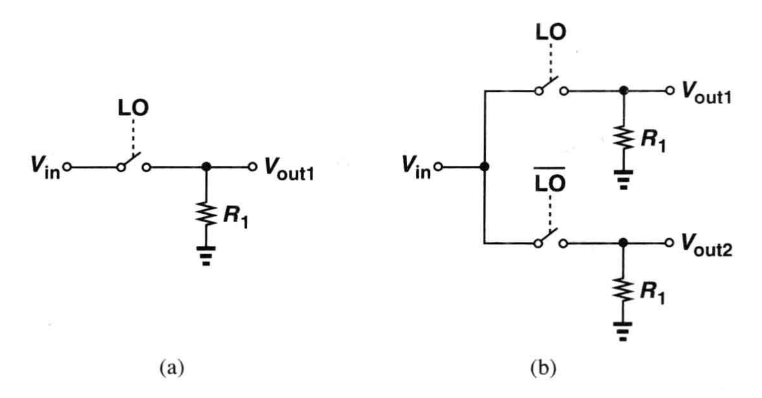


Figure 4.42 *(a) Simple mixer, (b) mixer with differential output.*

50% duty cycle, $S(t)$:

$$V_{out}(t) = V_{in}(t) \cdot S(t) \tag{4.25}$$

$$= V_{in}(t)\left[S(t) - \frac{1}{2}\right] + V_{in}(t) \cdot \frac{1}{2}. \tag{4.26}$$

We recognize that $S(t) - 1/2$ represents a "dc-free" square wave consisting of only odd harmonics. Thus, $V_{in}(t) \cdot [S(t) - 1/2]$ contains the product of V_{in} and the odd harmonics of the LO. The second term in (4.26), $V_{in}(t) \times 1/2$, denotes the RF feedthrough to the output (with no frequency translation).

Next, consider the topology depicted in Fig. 4.42(b), where a second branch driven by $\overline{\text{LO}}$ (the complement of LO) produces a second output. Expressing $\overline{\text{LO}}$ as $1 - S(t)$, we have

$$V_{out1}(t) = V_{in}(t)S(t) \tag{4.27}$$

$$V_{out2}(t) = V_{in}(t)[1 - S(t)]. \tag{4.28}$$

As with $V_{out1}(t)$, the second output $V_{out2}(t)$ contains an RF feedthrough equal to $V_{in}(t) \times 1/2$ because $1 - S(t)$ exhibits a dc content of $1/2$. If the output is sensed *differentially*, the RF feedthroughs in $V_{out1}(t)$ and $V_{out2}(t)$ are cancelled while the signal components add. It is this cancellation that is sensitive to asymmetries; for example, if the switches exhibit a mismatch between their on-resistances, then a net RF feedthrough arises in the differential output.

The problem of even-order distortion is critical enough to merit a quantitative measure. Called the "second intercept point" (IP_2), such a measure is defined according to a two-tone test similar to that for IP_3 except that the output of interest is the beat component rather than the intermodulation product. If $V_{in}(t) = A\cos\omega_1 t + A\cos\omega_2 t$, then the LNA output is given by

$$V_{out}(t) = \alpha_1 V_{in}(t) + \alpha_2 V_{in}^2(t) \tag{4.29}$$

$$= \alpha_1 A(\cos\omega_1 t + \cos\omega_2 t) + \alpha_2 A^2 \cos(\omega_1 + \omega_2)t$$
$$+ \alpha_2 A^2 \cos(\omega_1 - \omega_2)t + \cdots, \tag{4.30}$$

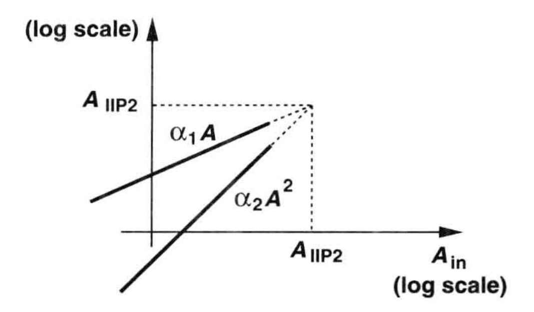


Figure 4.43 *Plot illustrating IP_2.*

Revealing that the beat amplitude grows with the *square* of the amplitude of the input tones. Thus, as shown in Fig. 4.43, the beat amplitude rises with a slope of 2 on a log scale. Since the net feedthrough of the beat depends on the mixer and LO asymmetries, the beat amplitude measured in the baseband depends on the device dimensions and the layout and is therefore difficult to formulate.

Example 4.20

Suppose the attenuation factor experienced by the beat as it travels through the mixer is equal to k, whereas the gain seen by each tone as it is downconverted to the baseband is equal to unity. Calculate the IP_2.

Solution:

From Eq. (4.30), the value of A that makes the output beat amplitude, $k\alpha_2 A^2$, equal to the main tone amplitude, $\alpha_1 A$, is given by

$$k\alpha_2 A_{IIP2}^2 = \alpha_1 A_{IIP2} \tag{4.31}$$

and hence

$$A_{IIP2} = \frac{1}{k} \cdot \frac{\alpha_1}{\alpha_2}. \tag{4.32}$$

Even-order distortion may manifest itself even in the absence of interferers. Suppose in addition to frequency and phase modulation, the received signal also exhibits *amplitude modulation*. For example, as explained in Chapter 3, QAM, OFDM, or simple QPSK with baseband pulse shaping produce variable-envelope waveforms. We express the signal as $x_{in}(t) = [A_0 + a(t)] \cos[\omega_c t + \phi(t)]$, where $a(t)$ denotes the envelope and typically varies slowly, i.e., it is a low-pass signal. Upon experiencing second-order distortion, the signal appears as

$$\alpha_2 x_{in}^2(t) = \alpha_2 \left[A_0^2 + 2A_0 a(t) + a^2(t)\right] \frac{1 + \cos[2\omega_c t + 2\phi(t)]}{2}. \tag{4.33}$$

Both of the terms $\alpha_2 A_0 a(t)$ and $\alpha_2 a^2(t)/2$ are *low-pass* signals and, like the beat component shown in Fig. 4.41, pass through the mixer with finite attenuation, corrupting the down-converted signal. We say even-order distortion demodulates AM because the amplitude information appears as $\alpha_2 A_0 a(t)$. This effect may corrupt the signal by its own envelope or by the envelope of a large interferer. We consider both cases below.

Example 4.21

Quantify the self-corruption expressed by Eq. (4.33) in terms of the IP_2.

Solution:

Assume, as in Example 4.20, that the low-pass components, $\alpha_2 A_0 a(t) + \alpha_2 a^2(t)/2$, experience an attenuation factor of k and the desired signal, $\alpha_1 A_0$, sees a gain of unity. Also, typically $a(t)$ is several times smaller than A_0 and hence the baseband corruption can be approximated as $k\alpha_2 A_0 a(t)$. Thus, the signal-to-noise ratio arising from self-corruption is given by

$$SNR = \frac{\alpha_1 A_0/\sqrt{2}}{k\alpha_2 A_0 a_{rms}} \tag{4.34}$$

$$= \frac{A_{IIP2}}{\sqrt{2} a_{rms}}, \tag{4.35}$$

where $A_0/\sqrt{2}$ denotes the rms signal amplitude and a_{rms} the rms value of $a(t)$.

How serious is the above phenomenon? Equation (4.35) predicts that the SNR falls to dangerously low levels as the envelope variation becomes comparable with the input IP_2. In reality, this is unlikely to occur. For example, if $a_{rms} = -20$ dBm, then A_0 is perhaps on the order of -10 to -15 dBm, large enough to saturate the receiver chain. For such high input levels, the gain of the LNA and perhaps the mixer is switched to much lower values to avoid saturation, automatically minimizing the above self-corruption effect.

The foregoing study nonetheless points to another, much more difficult, situation. If the desired channel is accompanied by a large amplitude-modulated interferer, then even-order distortion demodulates the AM component of the interferer, and mixer feedthrough allows it to appear in the baseband. In this case, Eq. (4.34) still applies but the numerator must represent the desired signal, $\alpha_1 A_{sig}/\sqrt{2}$, and the denominator, the interferer $k\alpha_2 A_{int} a_{rms}$:

$$SNR = \frac{\alpha_1 A_{sig}/\sqrt{2}}{k\alpha_2 A_{int} a_{rms}} \tag{4.36}$$

$$= \frac{A_{IIP2} A_{sig}/\sqrt{2}}{A_{int} a_{rms}}. \tag{4.37}$$

Example 4.22

A desired signal at $-100\,\text{dBm}$ is received along with an interferer $[A_{int} + a(t)]\cos[\omega_c t + \phi(t)]$, where $A_{int} = 5\text{ mV}$ and $a_{rms} = 1\text{ mV}$. What IP_2 is required to ensure SNR $\geq 20\,\text{dB}$?

Solution:

Since $-100\,\text{dBm}$ is equivalent to a peak amplitude of $A_{sig} = 3.16\,\mu\text{V}$, we have

$$A_{IIP2} = \text{SNR}\frac{A_{int}a_{rms}}{A_{sig}/\sqrt{2}} \tag{4.38}$$

$$= 22.4\,\text{V} \tag{4.39}$$

$$= +37\,\text{dBm}. \tag{4.40}$$

Note that the interferer level ($A_{int} = -36\,\text{dBm}$) falls well below the compression point of typical receivers, but it can still corrupt the signal if the IIP_2 is not as high as $+37\,\text{dBm}$.

This study reveals the relatively high IP_2 values required in direct-conversion receivers. We deal with methods of achieving a high IP_2 in Chapter 6.

Flicker Noise Since linearity requirements typically limit the gain of the LNA/mixer cascade to about 30 dB, the downconverted signal in a direct-conversion receiver is still relatively small and hence susceptible to noise in the baseband circuits. Furthermore, since the signal is centered around zero frequency, it can be substantially corrupted by flicker noise. As explained in Chapter 6, the mixers themselves may also generate flicker noise at their output.

In order to quantify the effect of flicker noise, let us assume the downconverted spectrum shown in Fig. 4.44, where f_{BW} is half of the RF channel bandwidth. The flicker noise is denoted by $S_{1/f}$ and the thermal noise at *the end* of the baseband by S_{th}. The frequency at which the two profiles meet is called f_c. We wish to determine the penalty due to flicker

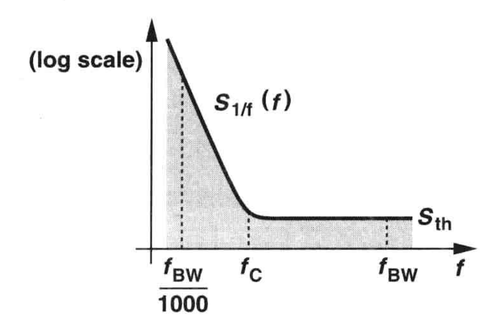


Figure 4.44 *Spectrum for calculation of flicker noise.*

noise, i.e., the additional noise power contributed by $S_{1/f}$. To this end, we note that if $S_{1/f} = \alpha/f$, then at f_c,

$$\frac{\alpha}{f_c} = S_{th}. \tag{4.41}$$

That is, $\alpha = f_c \cdot S_{th}$. Also, we assume noise components below roughly $f_{BW}/1000$ are unimportant because they vary so slowly that they negligibly affect the baseband symbols.[16] The total noise power from $f_{BW}/1000$ to f_{BW} is equal to

$$P_{n1} = \int_{f_{BW}/1000}^{f_c} \frac{\alpha}{f} df + (f_{BW} - f_c)S_{th} \tag{4.42}$$

$$= \alpha \ln \frac{1000 f_c}{f_{BW}} + (f_{BW} - f_c)S_{th} \tag{4.43}$$

$$= \left(6.9 + \ln \frac{f_c}{f_{BW}}\right) f_c S_{th} + (f_{BW} - f_c)S_{th} \tag{4.44}$$

$$= \left(5.9 + \ln \frac{f_c}{f_{BW}}\right) f_c S_{th} + f_{BW} S_{th}. \tag{4.45}$$

In the absence of flicker noise, the total noise power from $f_{BW}/1000$ to f_{BW} is given by

$$P_{n2} \approx f_{BW} S_{th}. \tag{4.46}$$

The ratio of P_{n1} and P_{n2} can serve as a measure of the flicker noise penalty:

$$\frac{P_{n1}}{P_{n2}} = 1 + \left(5.9 + \ln \frac{f_c}{f_{BW}}\right) \frac{f_c}{f_{BW}}. \tag{4.47}$$

Example 4.23

An 802.11g receiver exhibits a baseband flicker noise corner frequency of 200 kHz. Determine the flicker noise penalty.

Solution:

We have $f_{BW} = 10\,\text{MHz}, f_c = 200\,\text{kHz}$, and hence

$$\frac{P_{n1}}{P_{n2}} = 1.04. \tag{4.48}$$

How do the above results depend on the gain of the LNA/mixer cascade? In a good design, the thermal noise at the end of the baseband chain arises mostly from the noise of

16. As an extreme example, a noise component with a period of one day varies so slowly that it has negligible effect on a 20-minute phone conversation.

the antenna, the LNA, and the mixer. Thus, a higher front-end gain directly raises S_{th} in Fig. 4.44, thereby lowering the value of f_c and hence the flicker noise penalty.

Example 4.24

A GSM receiver exhibits a baseband flicker noise corner frequency of 200 kHz. Determine the flicker noise penalty.

Solution:

Figure 4.45 plots the baseband spectra, implying that the noise must be integrated up to

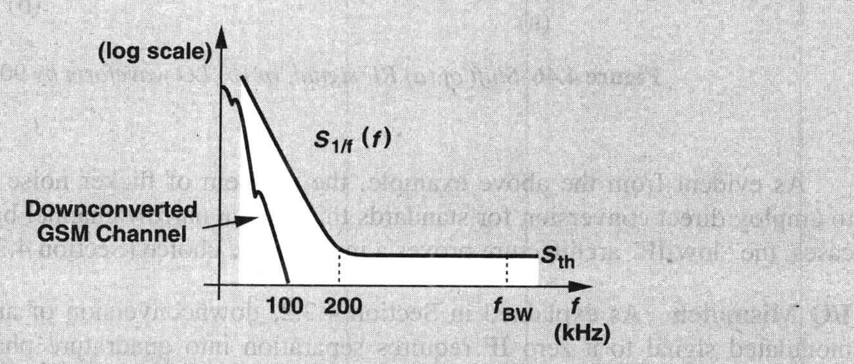


Figure 4.45 *Effect of flicker noise on a GSM channel.*

100 kHz. Assuming a lower end equal to about 1/1000 of the bit rate, we write the total noise as

$$P_{n1} = \int_{27\text{ Hz}}^{100\text{ kHz}} \frac{\alpha}{f} df \tag{4.49}$$

$$= f_c \cdot S_{th} \ln \frac{100\text{ kHz}}{27\text{ Hz}} \tag{4.50}$$

$$= 8.2 f_c S_{th}. \tag{4.51}$$

Without flicker noise,

$$P_{n2} \approx (100\text{ kHz}) S_{th}. \tag{4.52}$$

That is, the penalty reaches

$$\frac{P_{n1}}{P_{n2}} = \frac{8.2 f_c}{100\text{ kHz}} \tag{4.53}$$

$$= 16.4. \tag{4.54}$$

As expected, the penalty is much more severe in this case than in the 802.11g receiver of Example 4.23.

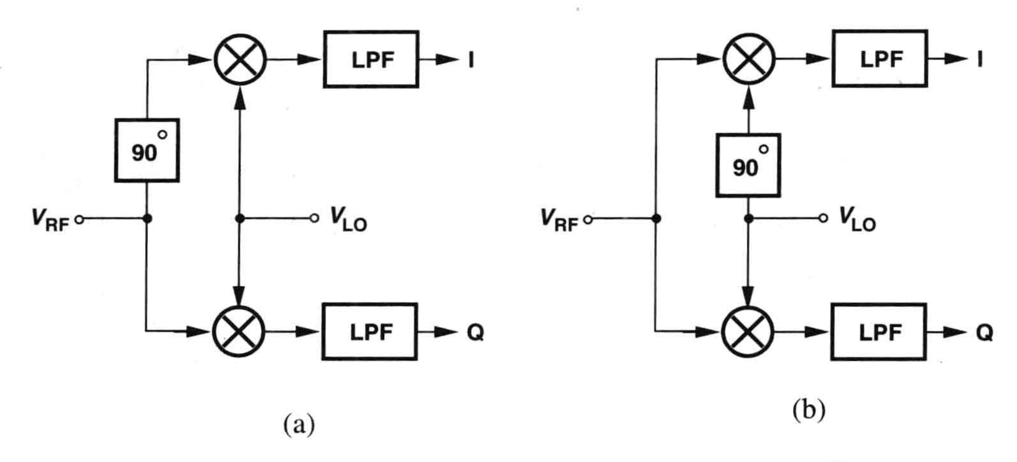


Figure 4.46 *Shift of (a) RF signal, or (b) LO waveform by* 90°.

As evident from the above example, the problem of flicker noise makes it difficult to employ direct conversion for standards that have a narrow channel bandwidth. In such cases, the "low-IF" architecture proves a more viable choice (Section 4.2.5).

I/Q Mismatch As explained in Section 4.2.2, downconversion of an asymmetrically-modulated signal to a zero IF requires separation into quadrature phases. This can be accomplished by shifting either the RF signal or the LO waveform by 90° (Fig. 4.46). Since shifting the RF signal generally entails severe noise-power-gain trade-offs, the approach in Fig. 4.46(b) is preferred. In either case, as illustrated in Fig. 4.47, errors in the 90° phase shift circuit and mismatches between the quadrature mixers result in imbalances in the amplitudes and phases of the baseband I and Q outputs. The baseband stages themselves may also contribute significant gain and phase mismatches.[17]

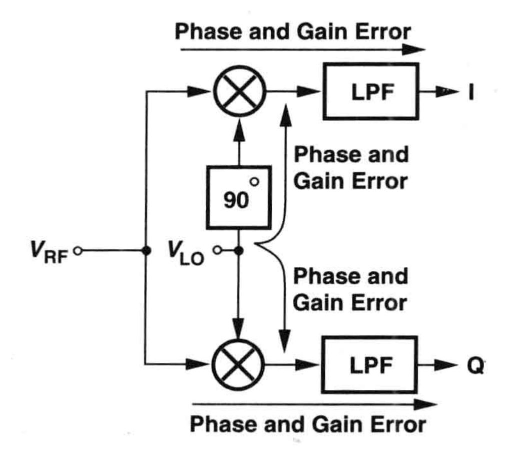


Figure 4.47 *Sources of I and Q mismatch.*

17. We use the terms "amplitude mismatch" and "gain mismatch" interchangeably.

Quadrature mismatches tend to be larger in direct-conversion receivers than in heterodyne topologies. This occurs because (1) the propagation of a higher frequency (f_{in}) through quadrature mixers experiences greater mismatches; for example, a delay mismatch of 10 ps between the two mixers translates to a phase mismatch of 18° at 5 GHz [Fig. 4.48(a)] and 3.6° at 1 GHz [Fig. 4.48(b)]; or (2) the quadrature phases of the LO itself suffer from greater mismatches at higher frequencies; for example, since device dimensions are reduced to achieve higher speeds, the mismatches between transistors increase.

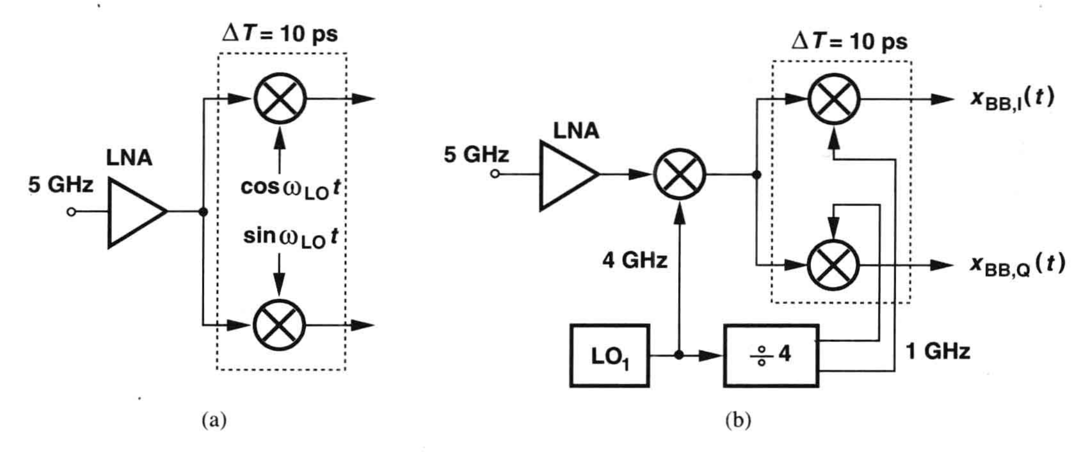


Figure 4.48 *Effect of a 10-ps propagation mismatch on a (a) direct-conversion and (b) heterodyne receiver.*

To gain insight into the effect of I/Q imbalance, consider a QPSK signal, $x_{in}(t) = a\cos\omega_c t + b\sin\omega_c t$, where a and b are either -1 or $+1$. Now let us lump all of the gain and phase mismatches shown in Fig. 4.47 in the LO path (Fig. 4.49)

$$x_{LO,I}(t) = 2\left(1+\frac{\epsilon}{2}\right)\cos\left(\omega_c t + \frac{\theta}{2}\right) \tag{4.55}$$

$$x_{LO,Q}(t) = 2\left(1-\frac{\epsilon}{2}\right)\sin\left(\omega_c t - \frac{\theta}{2}\right), \tag{4.56}$$

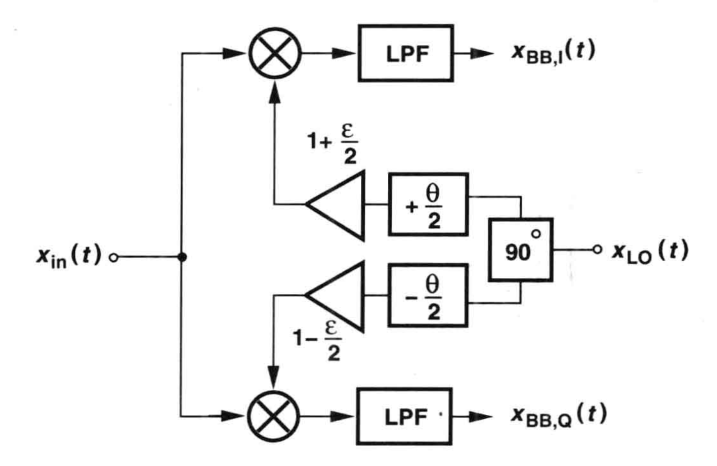


Figure 4.49 *Mismatches lumped in LO path.*

where the factor of 2 is included to simplify the results and ϵ and θ represent the amplitude and phase mismatches, respectively. Multiplying $x_{in}(t)$ by the quadrature LO waveforms and low-pass filtering the results, we obtain the following baseband signals:

$$x_{BB,I}(t) = a\left(1 + \frac{\epsilon}{2}\right)\cos\frac{\theta}{2} - b\left(1 + \frac{\epsilon}{2}\right)\sin\frac{\theta}{2} \tag{4.57}$$

$$x_{BB,Q}(t) = -a\left(1 - \frac{\epsilon}{2}\right)\sin\frac{\theta}{2} + b\left(1 - \frac{\epsilon}{2}\right)\cos\frac{\theta}{2}. \tag{4.58}$$

We now examine the results for two special cases: $\epsilon \neq 0$, $\theta = 0$ and $\epsilon = 0$, $\theta \neq 0$. In the former case, $x_{BB,I}(t) = a(1 + \epsilon/2)$ and $x_{BB,Q}(t) = b(1 - \epsilon/2)$, implying that the quadrature baseband symbols are scaled differently in amplitude [Fig. 4.50(a)]. More importantly, the points in the constellation are displaced [Fig. 4.50(b)].

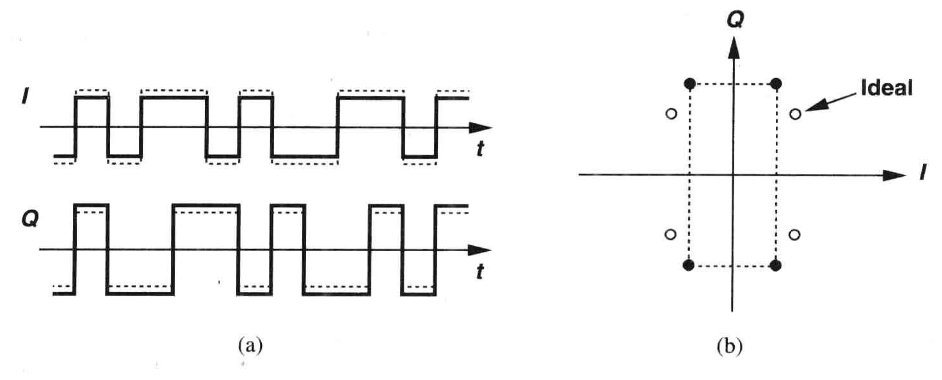


Figure 4.50 *Effect of gain mismatch on (a) time-domain waveforms and (b) constellation of a QPSK signal.*

With $\epsilon = 0\,\theta \neq 0$, we have $x_{BB,I}(t) = a\cos(\theta/2) - b\sin(\theta/2)$ and $x_{BB,Q}(t) = -a\sin(\theta/2) + b\cos(\theta/2)$. That is, each baseband output is corrupted by a fraction of the data symbols in the *other* output [Fig. 4.51(a)]. Also, the constellation is compressed along one diagonal and stretched along the other [Fig. 4.51(b)].

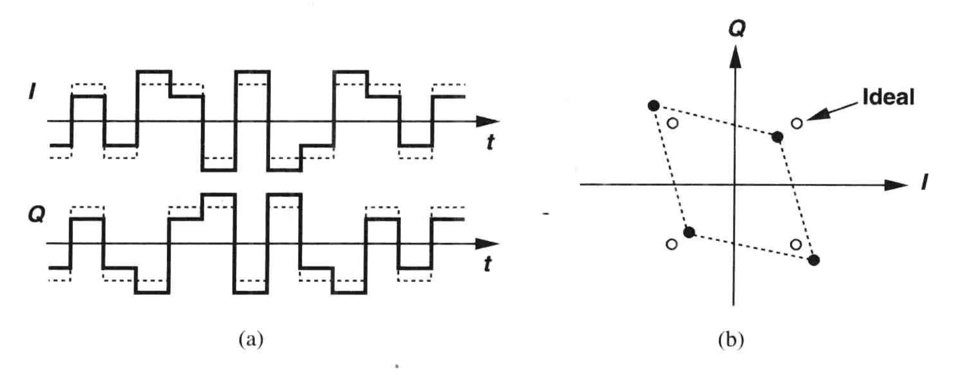


Figure 4.51 *Effect of phase mismatch on (a) time-domain waveforms and (b) constellation of a QPSK signal.*

Example 4.25

An FSK signal is applied to a direct-conversion receiver. Plot the baseband waveforms and determine the effect of I/Q mismatch.

Solution:

We express the FSK signal as $x_{FSK}(t) = A_0 \cos[(\omega_c + a\omega_1)t]$, where $a = \pm 1$ represents the binary information; i.e., the frequency of the carrier swings by $+\omega_1$ or $-\omega_1$. Upon multiplication by the quadrature phases of the LO, the signal produces the following baseband components:

$$x_{BB,I}(t) = -A_1 \cos a\omega_1 t \tag{4.59}$$

$$x_{BB,Q}(t) = +A_1 \sin a\omega_1 t. \tag{4.60}$$

Figure 4.52(a) illustrates the results: if the carrier frequency is equal to $\omega_c + \omega_1$ (i.e., $a = +1$), then the rising edges of $x_{BB,I}(t)$ coincide with the *positive* peaks of $x_{BB,Q}(t)$.

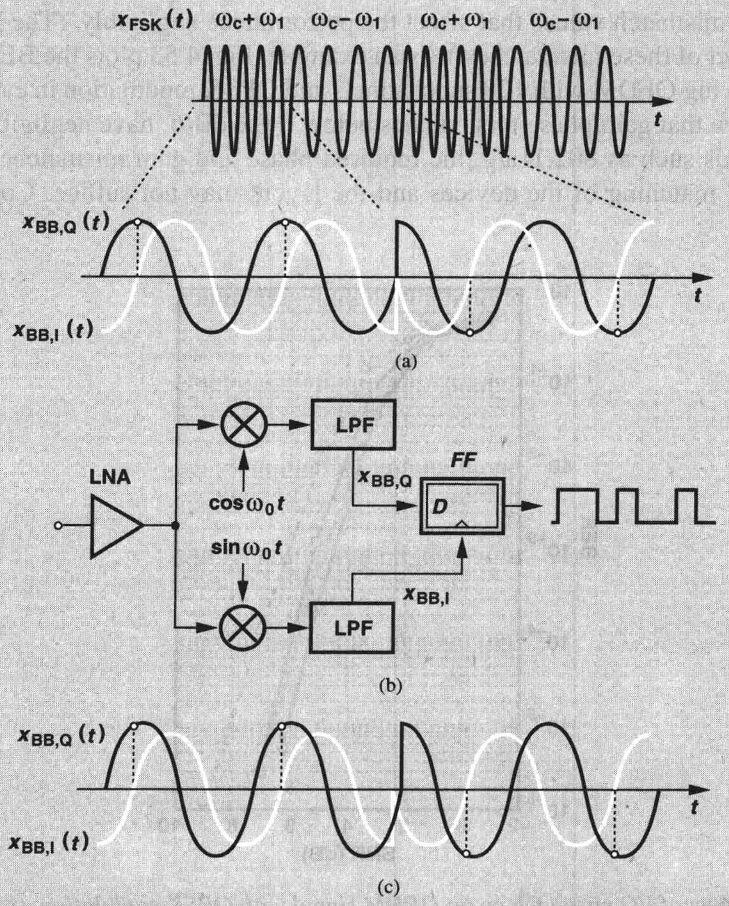


Figure 4.52 *(a) Baseband waveforms for an FSK signal, (b) FSK detection by a D flipflop, (c) effect of phase and gain mismatches.*

(*Continues*)

Example 4.25 *(Continued)*

Conversely, if the carrier frequency is equal to $\omega_c - \omega_1$, then the rising edges of $x_{BB,I}(t)$ coincide with the *negative* peaks of $x_{BB,Q}(t)$. Thus, the binary information is detected if $x_{BB,I}(t)$ simply *samples* $x_{BB,Q}(t)$, e.g., by means of a D flipflop [Fig. 4.52(b)].

The waveforms of Fig. 4.52(a) and the detection method of Fig. 4.52(b) suggest that FSK can tolerate large I/Q mismatches [Fig. 4.52(c)]: amplitude mismatch proves benign so long as the smaller output does not suffer from degraded SNR, and phase mismatch is tolerable so long as $x_{BB,I}(t)$ samples the correct *polarity* of $x_{BB,Q}(t)$. Of course, as the phase mismatch approaches 90°, the additive noise in the receive chain introduces errors.

In the design of an RF receiver, the maximum tolerable I/Q mismatch must be known so that the architecture and the building blocks are chosen accordingly. For complex signal waveforms such as OFDM with QAM, this maximum can be obtained by simulations: the bit error rate is plotted for different combinations of gain and phase mismatches, providing the maximum mismatch values that affect the performance negligibly. (The EVM can also reflect the effect of these mismatches.) As an example, Fig. 4.53 plots the BER curves for a system employing OFDM with 128 subchannels and QPSK modulation in each subchannel [1]. We observe that gain/phase mismatches below -0.6 dB/6° have negligible effect.

In standards such as 802.11a/g, the required phase and gain mismatches are so small that the "raw" matching of the devices and the layout may not suffice. Consequently, in

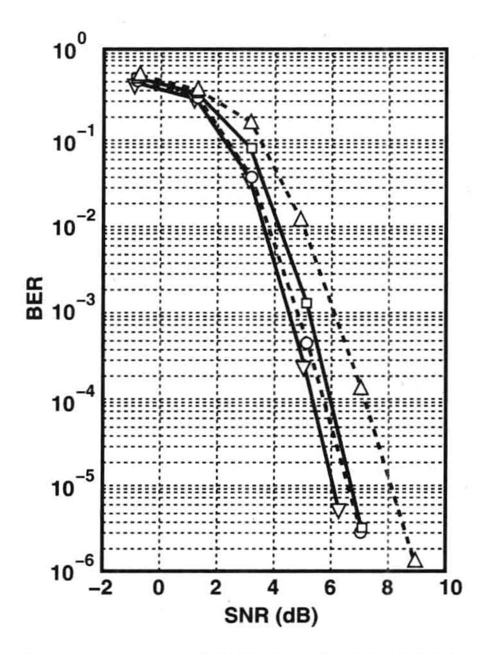


Figure 4.53 *Effect of I/Q mismatch on an OFDM signal with QPSK modulation. (▽: no imbalance; °: $\theta = 6^\circ, \epsilon = 0.6$ dB; □: $\theta = 10^\circ, \epsilon = 0.8$ dB; △: $\theta = 16^\circ, \epsilon = 1.4$ dB.)*

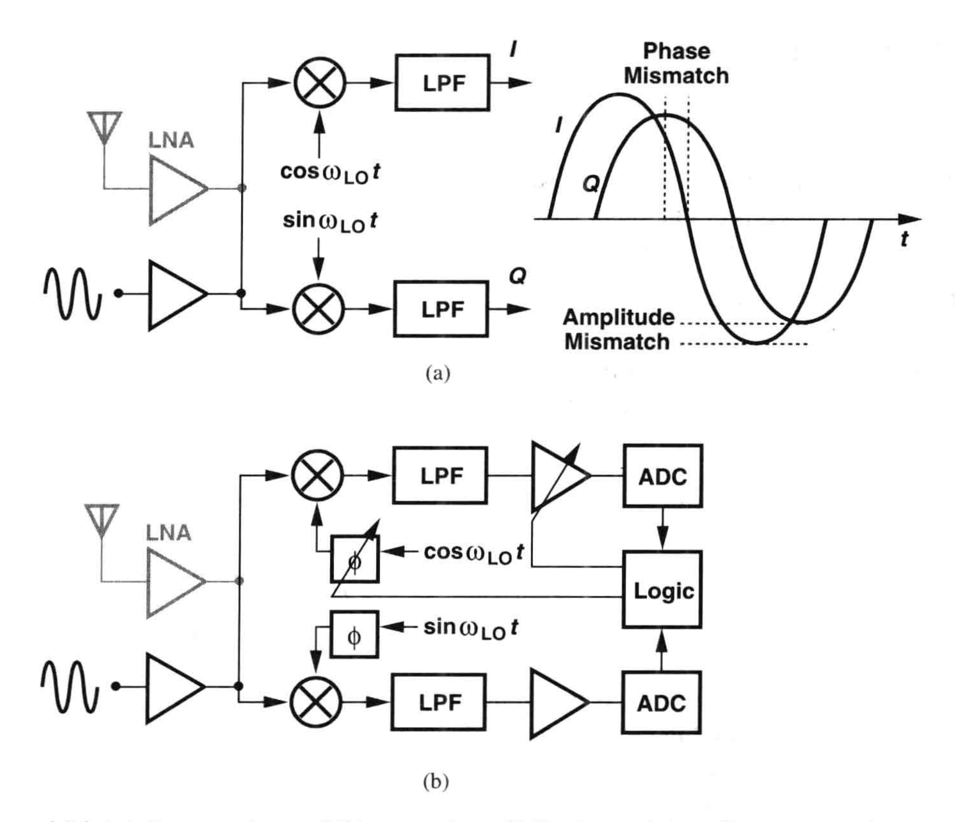


Figure 4.54 *(a) Computation and (b) correction of I/Q mismatch in a direct-conversion receiver.*

many high-performance systems, the quadrature phase and gain must be calibrated—either at power-up or continuously. As illustrated in Fig. 4.54(a), calibration at power-up can be performed by applying an RF tone at the input of the quadrature mixers and observing the baseband sinusoids in the analog or digital domain [2]. Since these sinusoids can be produced at arbitrarily low frequencies, their amplitude and phase mismatches can be measured accurately. With the mismatches known, the received signal constellation is corrected before detection. Alternatively, as depicted in Fig. 4.54(b), a variable-phase stage, ϕ, and a variable-gain stage can be inserted in the LO and baseband paths, respectively, and adjusted until the mismatches are sufficiently small. Note that the adjustment controls must be stored digitally during the actual operation of the receiver.

Mixing Spurs Unlike heterodyne systems, direct-conversion receivers rarely encounter corruption by mixing spurs. This is because, for an input frequency f_1 to fall in the baseband after experiencing mixing with nf_{LO}, we must have $f_1 \approx nf_{LO}$. Since f_{LO} is equal to the desired channel frequency, f_1 lies far from the band of interest and is greatly suppressed by the selectivity of the antenna, the band-select filter, and the LNA.

The issue of LO harmonics does manifest itself if the receiver is designed for a wide frequency band (greater than two octaves). Examples include TV tuners, "software-defined radios," and "cognitive radios."

4.2.4 Image-Reject Receivers

Our study of heterodyne and direct-conversion receivers has revealed various pros and cons. For example, heterodyning must deal with the image and mixing spurs and direct conversion, with even-order distortion and flicker noise. "Image-reject" architectures are another class of receivers that suppress the image without filtering, thereby avoiding the trade-off between image rejection and channel selection.

90° Phase Shift Before studying these architectures, we must define a "shift-by-90°" operation. First, let us consider a tone, $A\cos\omega_c t = (A/2)[\exp(+j\omega_c t) + \exp(-j\omega_c t)]$. The two exponentials respectively correspond to impulses at $+\omega_c$ and $-\omega_c$ in the frequency domain. We now shift the waveform by 90°:

$$A\cos(\omega_c t - 90^\circ) = A\frac{e^{+j(\omega_c t - 90^\circ)} + e^{-j(\omega_c t - 90^\circ)}}{2} \tag{4.61}$$

$$= -\frac{A}{2}je^{+j\omega_c t} + \frac{A}{2}je^{-j\omega_c t} \tag{4.62}$$

$$= A\sin\omega_c t. \tag{4.63}$$

Equivalently, the impulse at $+\omega_c$ is multiplied by $-j$ and that at $-\omega_c$, by $+j$. We illustrate this transformation in the three-dimensional diagram of Fig. 4.55(a), recognizing that the impulse at $+\omega_c$ is rotated clockwise and that at $-\omega_c$ counterclockwise.

Similarly, for a narrowband modulated signal, $x(t) = A(t)\cos[\omega_c t + \phi(t)]$, we perform a 90° phase shift as

$$A(t)\cos[\omega_c t + \phi(t) - 90^\circ] = A(t)\frac{e^{+j[\omega_c t + \phi(t) - 90^\circ]} + e^{-j[\omega_c t + \phi(t) - 90^\circ]}}{2} \tag{4.64}$$

$$= A(t)\frac{-je^{+j[\omega_c t + \phi(t)]} + je^{-j[\omega_c t + \phi(t)]}}{2} \tag{4.65}$$

$$= A(t)\sin[\omega_c t + \phi(t)]. \tag{4.66}$$

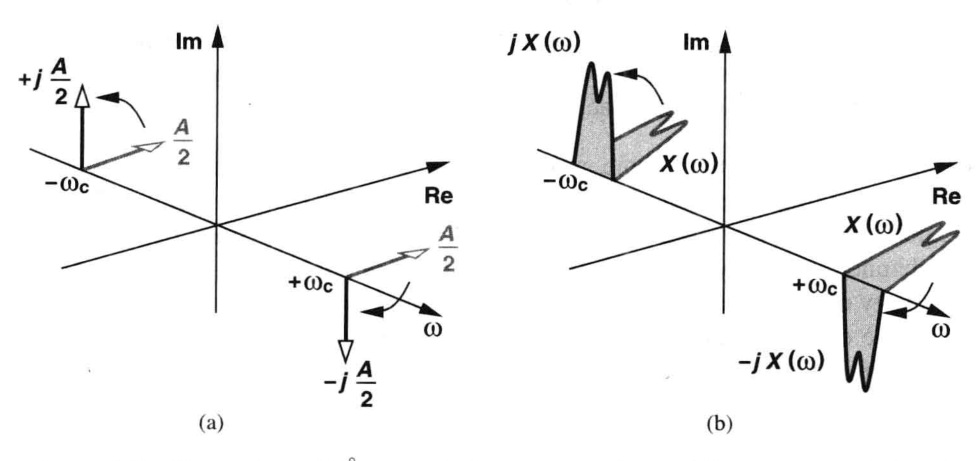


Figure 4.55 *Illustration of 90° phase shift for (a) a cosine and (b) a modulated signal.*

As depicted in Fig. 4.55(b), the positive-frequency contents are multiplied by $-j$ and the negative-frequency contents by $+j$ (if ω_c is positive). Alternatively, we write in the frequency domain:

$$X_{90^\circ}(\omega) = X(\omega)[-j\mathrm{sgn}(\omega)], \tag{4.67}$$

where $\mathrm{sgn}(\omega)$ denotes the signum (sign) function. The shift-by-90° operation is also called the "Hilbert transform." The reader can prove that the Hilbert transform of the Hilbert transform (i.e., the cascade of two 90° phase shifts) simply negates the original signal.

Example 4.26

In phasor diagrams, we simply multiply a phasor by $-j$ to rotate it by 90° clockwise. Is that inconsistent with the Hilbert transform?

Solution:

No, it is not. A phasor is a representation of $A\exp(j\omega_c t)$, i.e., only the positive frequency content. That is, we implicitly assume that if $A\exp(j\omega_c t)$ is multiplied by $-j$, then $A\exp(-j\omega_c t)$ is also multiplied by $+j$.

The Hilbert transform, as expressed by Eq. (4.67), *distinguishes* between negative and positive frequencies. This distinction is the key to image rejection.

Example 4.27

Plot the spectrum of $A\cos\omega_c t + jA\sin\omega_c t$.

Solution:

Multiplication of the spectrum of $A\sin\omega_c t$ by j rotates both impulses by 90° counterclockwise [Fig. 4.56(a)]. Upon adding this spectrum to that $A\cos\omega_c t$, we obtain the

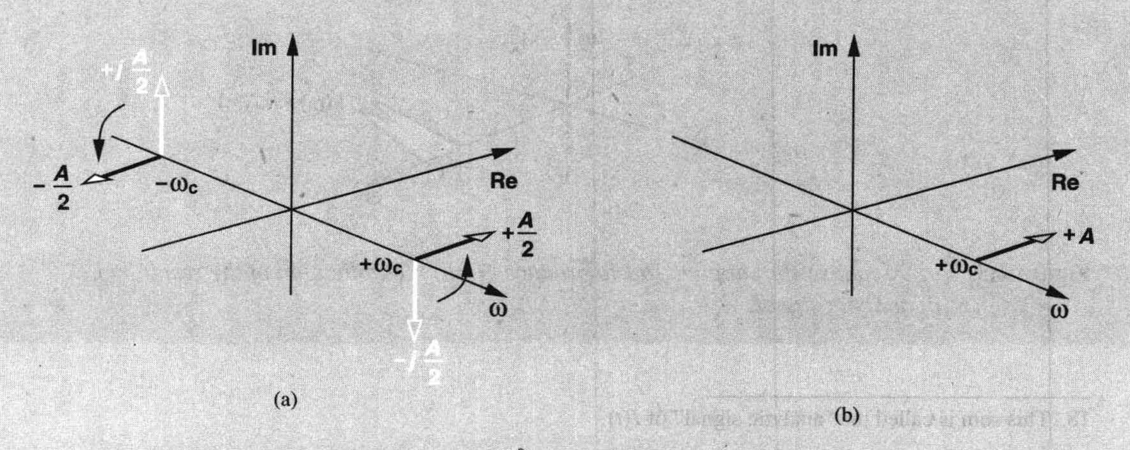


Figure 4.56 *(a) A sine subjected to 90° phase shift, (b) spectrum of $A\cos\omega_c t + j\sin\omega_c t$.*

(Continues)

Example 4.27 *(Continued)*

one-sided spectrum shown in Fig. 4.56(b). This is, of course, to be expected because $A\cos\omega_c t + jA\sin\omega_c t = A\exp(-j\omega_c t)$, whose Fourier transform is a single impulse located at $\omega = +\omega_c$.

Example 4.28

A narrowband signal $I(t)$ with a real spectrum is shifted by 90° to produce $Q(t)$. Plot the spectrum of $I(t) + jQ(t)$.[18]

Solution:

We first multiply $I(\omega)$ by $-j\text{sgn}(\omega)$ [Fig. 4.57(a)] and then, in a manner similar to the previous example, multiply the result by j [Fig. 4.57(b)]. The spectrum of $jQ(t)$ therefore cancels that of $I(t)$ at negative frequencies and enhances it at positive frequencies [Fig. 4.57(c)]. The one-sided spectrum of $I(t) + jQ(t)$ proves useful in the analysis of transceivers.

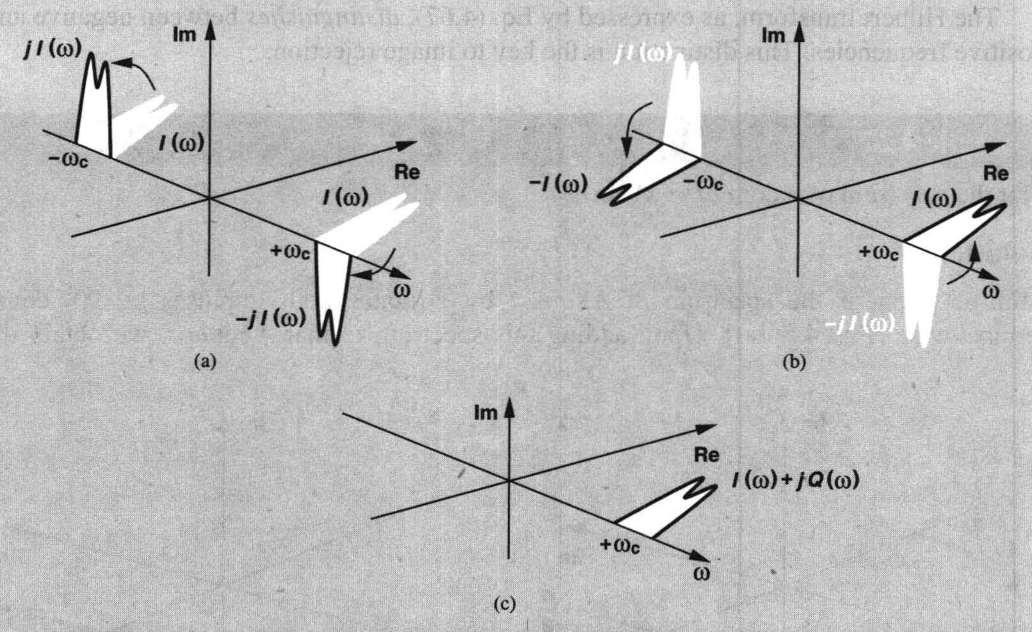
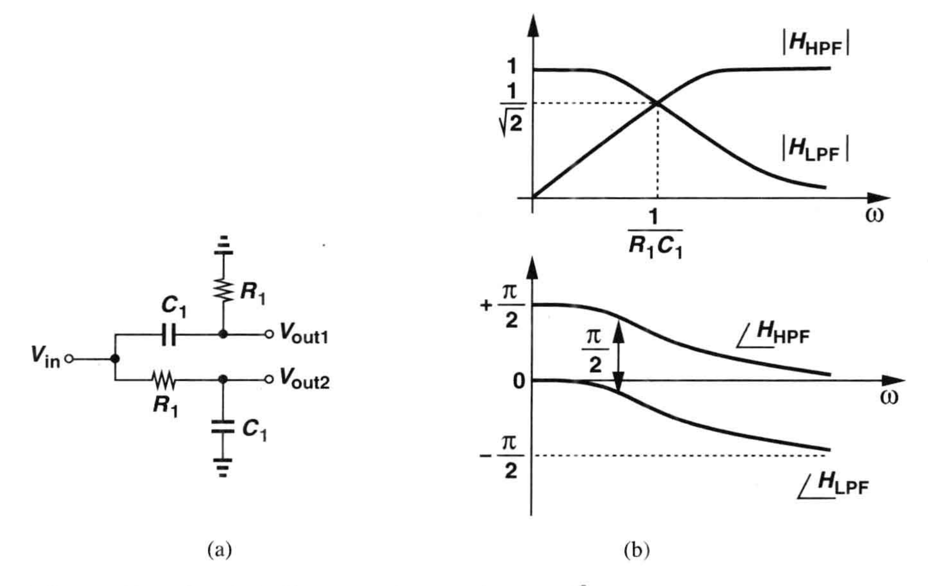


Figure 4.57 *(a) 90° phase shift applied to I to produce Q, (b) multiplication of the result by j, (c) analytic signal.*

18. This sum is called the "analytic signal" of $I(t)$.



(a) (b)

Figure 4.58 *(a) Use of an RC-CR network to perform a 90° phase shift, (b) frequency response of the network.*

How is the 90° phase shift implemented? Consider the *RC-CR* network shown in Fig. 4.58(a), where the high-pass and low-pass transfer functions are respectively given by

$$H_{HPF}(s) = \frac{V_{out1}}{V_{in}} = \frac{R_1C_1s}{R_1C_1s+1} \quad (4.68)$$

$$H_{LPF}(s) = \frac{V_{out2}}{V_{in}} = \frac{1}{R_1C_1s+1}. \quad (4.69)$$

The transfer functions exhibit a phase of $\angle H_{HPF} = \pi/2 - \tan^{-1}(R_1C_1\omega)$ and $\angle H_{LPF} = -\tan^{-1}(R_1C_1\omega)$. Thus, $\angle H_{HPF} - \angle H_{LPF} = \pi/2$ at all frequencies and any choice of R_1 and C_1. Also, $|V_{out1}/V_{in}| = |V_{out2}/V_{in}| = 1/\sqrt{2}$ at $\omega = (R_1C_1)^{-1}$ [Fig. 4.58(b)]. We can therefore consider V_{out2} as the Hilbert transform of V_{out1} at frequencies close to $(R_1C_1)^{-1}$.

Another approach to realizing the 90°-phase-shift operation is illustrated in Fig. 4.59(a), where the RF input is mixed with the quadrature phases of the LO so as to translate the spectrum to a nonzero IF. As shown in Fig. 4.59(b), as a result of mixing with $\cos\omega_{LO}t$, the impulse at $-\omega_{LO}$ is convolved with the input spectrum around $+\omega_c$, generating that at $-\omega_{IF}$. Similarly, the impulse at $+\omega_{LO}$ produces the spectrum at $+\omega_{IF}$ from that at $-\omega_c$. Depicted in Fig. 4.59(c), mixing with $\sin\omega_{LO}t$ results in an IF spectrum at $-\omega_{IF}$ with a coefficient $+j/2$ and another at $+\omega_{IF}$ with a coefficient $-j/2$. We observe that, indeed, the IF spectrum emerging from the lower arm is the Hilbert transform of that from the upper arm.

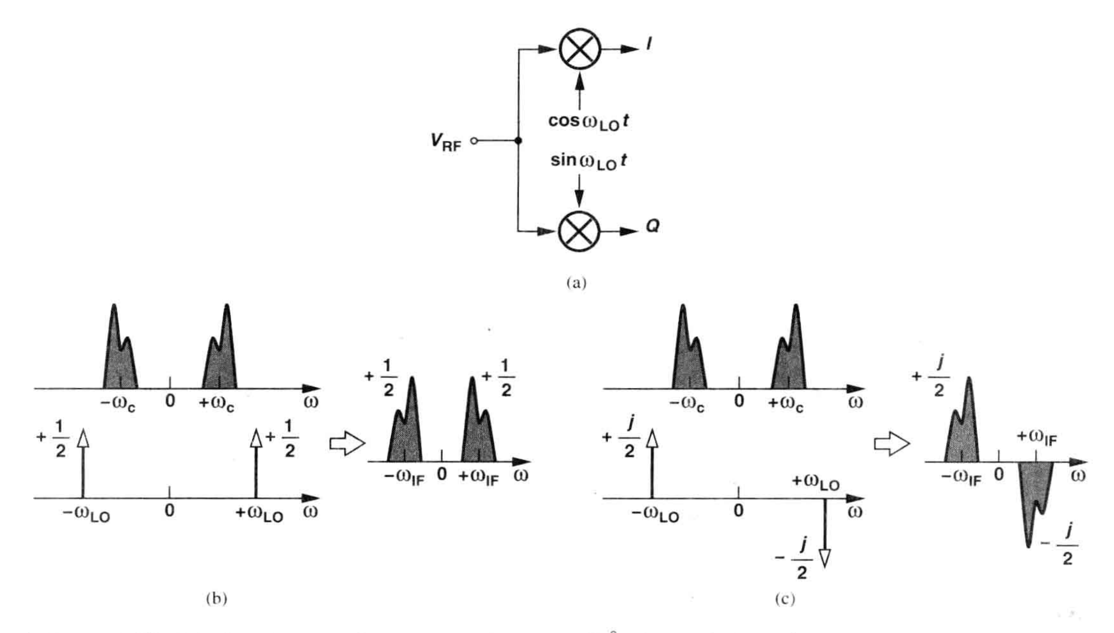


Figure 4.59 *(a) Quadrature downconversion as a 90° phase shifter, (b) output spectrum resulting from multiplication by* $\cos\omega_{LO}t$*, (c) output spectrum resulting from multiplication by* $\sin\omega_{LO}t$.

Example 4.29

The realization of Fig. 4.59(a) assumes high-side injection for the LO. Repeat the analysis for low-side injection.

Solution:

Figures 4.60(a) and (b) show the spectra for mixing with $\cos\omega_{LO}t$ and $\sin\omega_{LO}t$, respectively. In this case, the IF component in the lower arm is the *negative* of the Hilbert transform of that in the upper arm.

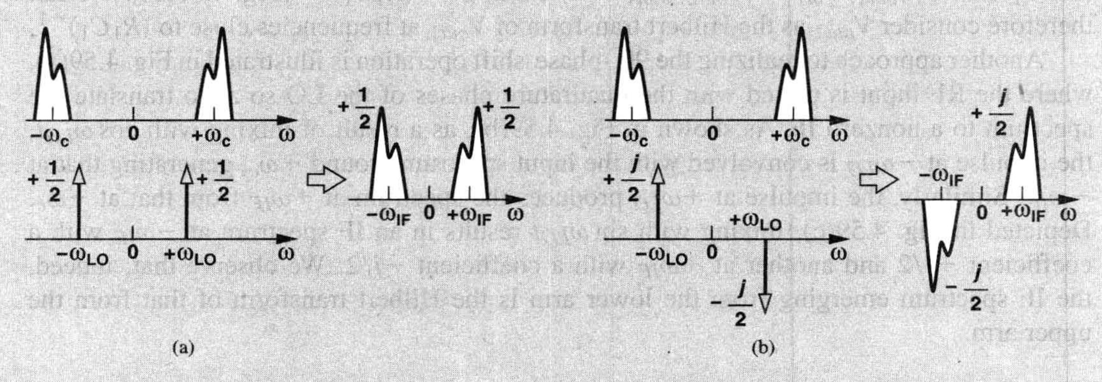


Figure 4.60 *Low-side-injection mixing of an RF signal with (a)* $\cos\omega_{LO}t$ *and (b)* $\sin\omega_{LO}t$.

Let us summarize our findings thus far. The quadrature converter[19] of Fig. 4.59(a) produces at its output a signal and its Hilbert transform if $\omega_c > \omega_{LO}$ or a signal and the negative of its Hilbert transform if $\omega_c < \omega_{LO}$. This arrangement therefore distinguishes between the desired signal and its image. Figure 4.61 depicts the three-dimensional IF spectra if a signal and its image are applied at the input and $\omega_{LO} < \omega_c$.

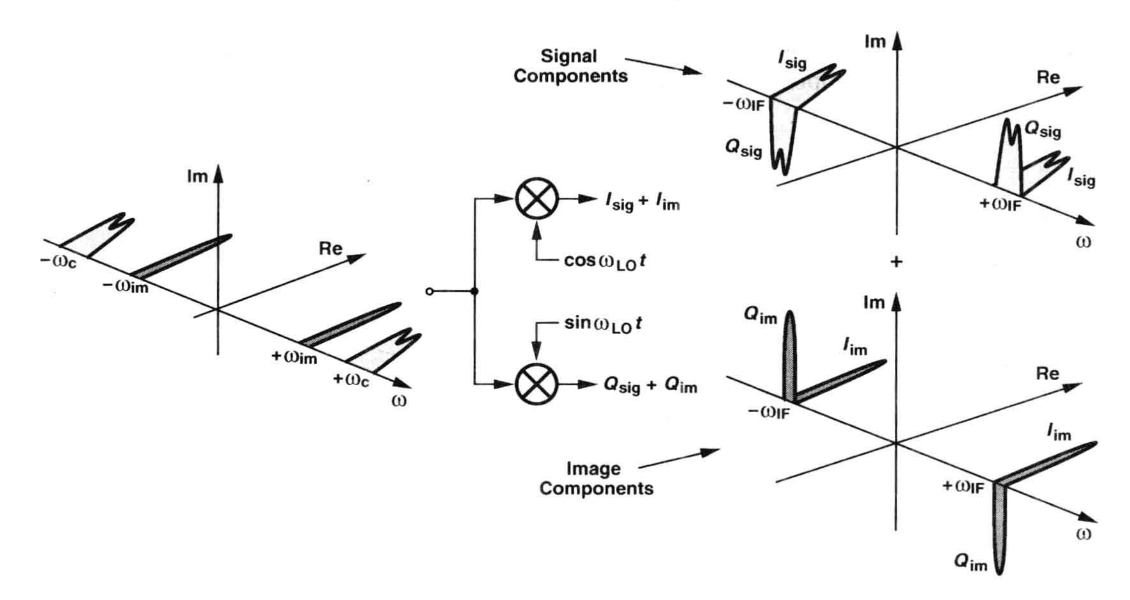


Figure 4.61 *Input and output spectra in a quadrature downconverter with low-side injection.*

Hartley Architecture How can the image components in Fig. 4.61 cancel each other? For example, is $I(t) + Q(t)$ free from the image? Since the image components in $Q(t)$ are 90° out of phase with respect to those in $I(t)$, this summation still contains the image. However, since the Hilbert transform of the Hilbert transform negates the signal, if we shift $I(t)$ or $Q(t)$ by another 90° before adding them, the image may be removed. This hypothesis forms the foundation for the Hartley architecture shown in Fig. 4.62. (The original idea proposed

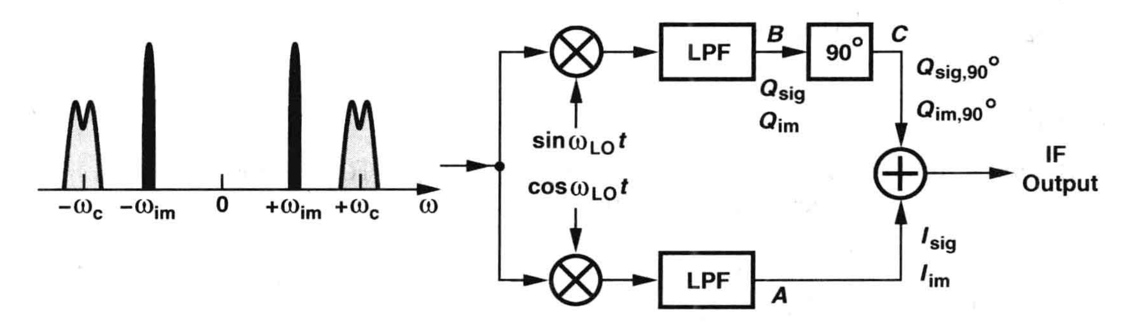


Figure 4.62 *Hartley image-reject receiver.*

19. We can also consider this a quadrature *down*converter if $\omega_{IF} < \omega_c$.

by Hartley relates to single-sideband transmitters [4].) The low-pass filters are inserted to remove the unwanted high-frequency components generated by the mixers.

To understand the operation of Hartley's architecture, we assume low-side injection and apply a 90° phase shift to the Hilbert transforms of the signal and the image (the Q arm) in Fig. 4.61, obtaining $Q_{sig,90^\circ}$ and $Q_{im,90^\circ}$ as shown in Fig. 4.63. Multiplication of Q_{sig} by $-j\text{sgn}(\omega)$ rotates and superimposes the spectrum of Q_{sig} on that of I_{sig} (from Fig. 4.61), doubling the signal amplitude. On the other hand, multiplication of Q_{im} by $-j\text{sgn}(\omega)$ creates the opposite of I_{im}, cancelling the image.

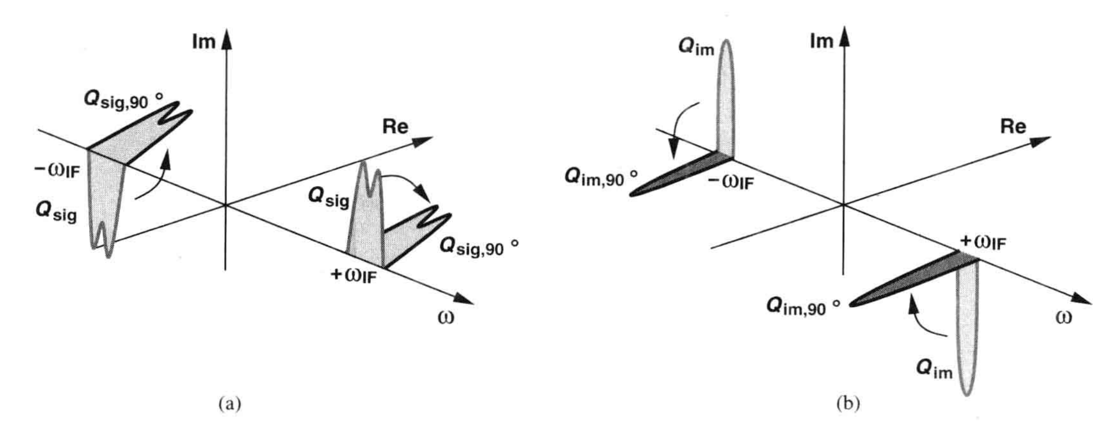


Figure 4.63 *Spectra at points B and C in Hartley receiver.*

In summary, the Hartley architecture first takes the negative Hilbert transform of the signal and the Hilbert transform of the image (or vice versa) by means of quadrature mixing, subsequently takes the Hilbert transform of one of the downconverted outputs, and sums the results. That is, the signal spectrum is multiplied by $[+j\text{sgn}(\omega)][-j\text{sgn}(\omega)] = +1$, whereas the image spectrum is multiplied by $[-j\text{sgn}(\omega)][-j\text{sgn}(\omega)] = -1$.

Example 4.30

An eager student constructs the Hartley architecture but with high-side injection. Explain what happens.

Solution:

From Fig. 4.60, we note that the quadrature converter takes the Hilbert transform of the signal and the negative Hilbert transform of the image. Thus, with another 90° phase shift, the outputs C and A in Fig. 4.62 contain the signal with *opposite* polarities and the image with the same polarity. The circuit therefore operates as a "signal-reject" receiver! Of course, the design is salvaged if the addition is replaced with subtraction.

The behavior of the Hartley architecture can also be expressed analytically. Let us represent the received signal and image as $x(t) = A_{sig}\cos(\omega_c t + \phi_{sig}) + A_{im}\cos(\omega_{im} t + \phi_{im})$, where the amplitudes and phases are functions of time in the general case. Multiplying $x(t)$ by the LO phases and neglecting the high-frequency components, we obtain the signals at

points A and B in Fig. 4.62:

$$x_A(t) = \frac{A_{sig}}{2}\cos[(\omega_c - \omega_{LO})t + \phi_{sig}] + \frac{A_{im}}{2}\cos[(\omega_{im} - \omega_{LO})t + \phi_{im}] \quad (4.70)$$

$$x_B(t) = -\frac{A_{sig}}{2}\sin[(\omega_c - \omega_{LO})t + \phi_{sig}] - \frac{A_{im}}{2}\sin[(\omega_{im} - \omega_{LO})t + \phi_{im}], \quad (4.71)$$

where a unity LO amplitude is assumed for simplicity. Now, $x_B(t)$ must be shifted by 90°. With low-side injection, the first sine has a positive frequency and becomes negative of a cosine after the 90° shift (why?). The second sine, on the other hand, has a negative frequency. We therefore write $-(A_{im}/2)\sin[(\omega_{im} - \omega_{LO})t + \phi_{im}] = (A_{im}/2)\sin[(\omega_{LO} - \omega_{im})t - \phi_{im}]$ so as to obtain a positive frequency and shift the result by 90°, arriving at $-(A_{im}/2)\cos[(\omega_{LO} - \omega_{im})t - \phi_{im}] = -(A_{im}/2)\cos[(\omega_{im} - \omega_{LO})t + \phi_{im}]$. It follows that

$$x_C(t) = \frac{A_{sig}}{2}\cos[(\omega_c - \omega_{LO})t + \phi_{sig}] - \frac{A_{im}}{2}\cos[(\omega_{im} - \omega_{LO})t + \phi_{im}]. \quad (4.72)$$

Upon addition of $x_A(t)$ and $x_C(t)$, we retain the signal and reject the image.

The 90° phase shift depicted in Fig. 4.62 is typically realized as a $+45^\circ$ shift in one path and -45° shift in the other (Fig. 4.64). This is because it is difficult to shift a single signal by 90° while circuit components vary with process and temperature.

The principal drawback of the Hartley architecture stems from its sensitivity to mismatches: the perfect image cancellation described above occurs only if the amplitude and phase of the negative of the image exactly match those of the image itself. If the LO phases are not in exact quadrature or the gains and phase shifts of the upper and lower arms in Fig. 4.64 are not identical, then a fraction of the image remains. To quantify this effect, we lump the mismatches of the receiver as a single amplitude error, ϵ, and phase error, $\Delta\theta$, in the LO path, i.e., one LO waveform is expressed as $\sin\omega_{LO}t$ and the other as $(1+\epsilon)\cos(\omega_{LO}t + \Delta\theta)$. Expressing the received signal and image as $x(t) = A_{sig}\cos(\omega_c t + \phi_{sig}) + A_{im}\cos(\omega_{im}t + \phi_{im})$ and multiplying $x(t)$ by the LO waveforms, we write the downconverted signal at point A in Fig. 4.62 as

$$x_A(t) = \frac{A_{sig}}{2}(1+\epsilon)\cos[(\omega_c - \omega_{LO})t + \phi_{sig} + \Delta\theta] + \frac{A_{im}}{2}(1+\epsilon)\cos[(\omega_{im} - \omega_{LO})t + \phi_{im} + \Delta\theta]. \quad (4.73)$$

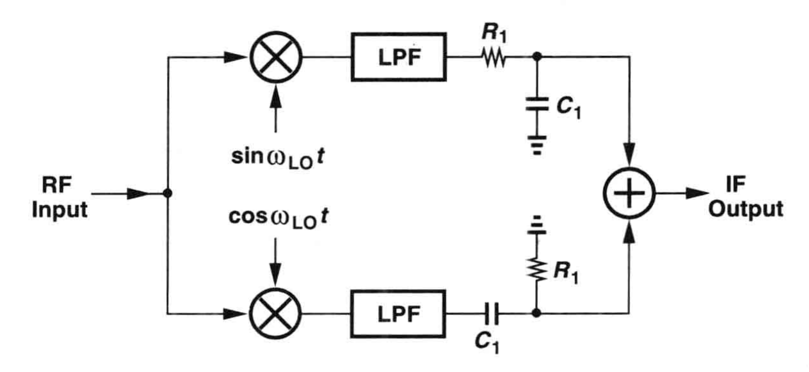


Figure 4.64 *Realization of 90° phase shift in Hartley receiver.*

The spectra at points B and C are still given by Eqs. (4.71) and (4.72), respectively. We now add $x_A(t)$ and $x_C(t)$ and group the signal and image components at the output:

$$x_{sig}(t) = \frac{A_{sig}}{2}(1+\epsilon)\cos[(\omega_c - \omega_{LO})t + \phi_{sig} + \Delta\theta] + \frac{A_{sig}}{2}\cos[(\omega_c - \omega_{LO})t + \phi_{sig}] \quad (4.74)$$

$$x_{im}(t) = \frac{A_{im}}{2}(1+\epsilon)\cos[(\omega_{im} - \omega_{LO})t + \phi_{im} + \Delta\theta] - \frac{A_{im}}{2}\cos[(\omega_{im} - \omega_{LO})t + \phi_{im}]. \quad (4.75)$$

To arrive at a meaningful measure of the image rejection, we divide the image-to-signal ratio at the input by the same ratio at the output.[20] The result is called the "image rejection ratio" (IRR). Noting that the average power of the vector sum $a\cos(\omega t + \alpha) + b\cos\omega t$ is given by $(a^2 + 2ab\cos\alpha + b^2)/2$, we write the output image-to-signal ratio as

$$\frac{P_{im}}{P_{sig}}\bigg|_{out} = \frac{A_{im}^2}{A_{sig}^2}\frac{(1+\epsilon)^2 - 2(1+\epsilon)\cos\Delta\theta + 1}{(1+\epsilon)^2 + 2(1+\epsilon)\cos\Delta\theta + 1}. \quad (4.76)$$

Since the image-to-signal ratio at the input is given by A_{im}^2/A_{sig}^2, the IRR can be expressed as

$$\text{IRR} = \frac{(1+\epsilon)^2 + 2(1+\epsilon)\cos\Delta\theta + 1}{(1+\epsilon)^2 - 2(1+\epsilon)\cos\Delta\theta + 1}. \quad (4.77)$$

Note that ϵ denotes the *relative* gain error and $\Delta\theta$ is in radians. Also, to express IRR in dB, we must compute 10 log IRR (rather than 20 log IRR).

Example 4.31

If $\epsilon \ll 1$ rad, simplify the expression for IRR.

Solution:

Since $\cos\Delta\theta \approx 1 - \Delta\theta^2/2$ for $\Delta\theta \ll 1$ rad, we can reduce (4.77) to

$$\text{IRR} \approx \frac{4 + 4\epsilon + \epsilon^2 - (1+\epsilon)\Delta\theta^2}{\epsilon^2 + (1+\epsilon)\Delta\theta^2}. \quad (4.78)$$

In the numerator, the first term dominates and in the denominator $\epsilon \ll 1$, yielding

$$\text{IRR} \approx \frac{4}{\epsilon^2 + \Delta\theta^2}. \quad (4.79)$$

20. Note that the ratio of the output image power and the input image power is not meaningful because it depends on the gain.

Example 4.31 *(Continued)*

For example, $\epsilon = 10\%$ (≈ 0.83 dB)[21] limits the IRR to 26 dB. Similarly, $\Delta\theta = 10^{\circ}$ yields an IRR of 21 dB. While such mismatch values may be tolerable in direct-conversion receivers, they prove inadequate here.

With various mismatches arising in the LO and signal paths, the IRR typically falls below roughly 35 dB. This issue and a number of other drawbacks limit the utility of the Hartley architecture.

Another critical drawback, especially in CMOS technology, originates from the variation of the absolute values of R_1 and C_1 in Fig. 4.64. Recall from Fig. 4.58 that the phase shift produced by the *RC-CR* network remains equal to 90° even with such variations, but the output amplitudes are equal at only $\omega = (R_1C_1)^{-1}$. Specifically, if R_1 and C_1 are nominally chosen for a certain IF, $(R_1C_1)^{-1} = \omega_{IF}$, but respectively experience a small change of ΔR and ΔC with process or temperature, then the ratio of the output amplitudes of the high-pass and low-pass sections is given by

$$\left|\frac{H_{HPF}}{H_{LPF}}\right| = (R_1 + \Delta R)(C_1 + \Delta C)\omega_{IF} \tag{4.80}$$

$$\approx 1 + \frac{\Delta R}{R_1} + \frac{\Delta C}{C_1}. \tag{4.81}$$

Thus, the gain mismatch is equal to

$$\epsilon = \frac{\Delta R}{R_1} + \frac{\Delta C}{C_1}. \tag{4.82}$$

For example, $\Delta R/R_1 = 20\%$ limits the image rejection to only 20 dB. Note that these calculations have assumed perfect matching between the high-pass and low-pass sections. If the resistors or capacitors exhibit mismatches, the IRR degrades further.

Another drawback resulting from the *RC-CR* sections manifests itself if the signal translated to the IF has a wide bandwidth. Since the gains of the high-pass and low-pass sections depart from each other as the frequency departs from $\omega_{IF} = (R_1C_1)^{-1}$ [Fig. 4.58(b)], the image rejection may degrade substantially near the edges of the channel. In Problem 4.16, the reader can prove that, at a frequency of $\omega_{IF} + \Delta\omega$, the IRR is given by

$$\text{IRR} = \left(\frac{\omega_{IF}}{\Delta\omega}\right)^2. \tag{4.83}$$

For example, a fractional bandwidth of $2\Delta\omega/\omega_{IF} = 5\%$ limits the IRR to 32 dB.

The limitation expressed by Eq. (4.83) implies that ω_{IF} cannot be zero, dictating a heterodyne approach. Figure 4.65 shows an example where the first IF is followed by another quadrature downconverter so as to produce the baseband signals. Unlike the sliding-IF

21. To calculate ϵ in dB, we write $20\log(1 + 10\%) = 0.83$ dB.

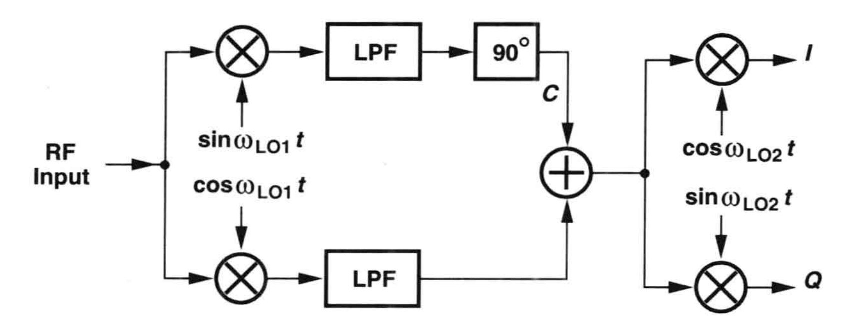


Figure 4.65 *Downconversion of Hartley receiver output to baseband.*

architecture of Fig. 4.26(a), this topology also requires the quadrature phases of the first LO, a critical disadvantage. The mixing spurs studied in Section 4.2.1 persist here as well.

The *RC-CR* sections used in Fig. 4.64 also introduce attenuation and noise. The 3-dB loss resulting from $|H_{HPF}| = |H_{LPF}| = 1/\sqrt{2}$ at $\omega = (R_1C_1)^{-1}$ directly amplifies the noise of the following adder. Moreover, the input impedance of each section, $|R_1 + (C_1s)^{-1}|$, reaches $\sqrt{2}R_1$ at $\omega = (R_1C_1)^{-1}$, imposing a trade-off between the loading seen by the mixers and the thermal noise of the 90° shift circuit.

The voltage adder at the output of the Hartley architecture also poses difficulties as its noise and nonlinearity appear in the signal path. Illustrated in Fig. 4.66, the summation is typically realized by differential pairs, which convert the signal voltages to currents, sum the currents, and convert the result to a voltage.

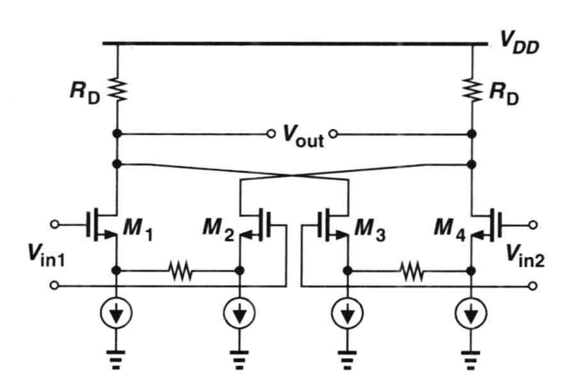


Figure 4.66 *Summation of two voltages.*

Weaver Architecture Our analysis of the Hartley architecture has revealed several issues that arise from the use of the *RC-CR* phase shift network. The Weaver receiver, derived from its transmitter counterpart [5], avoids these issues.

As recognized in Fig. 4.59, mixing a signal with quadrature phases of an LO takes the Hilbert transform. Depicted in Fig. 4.67, the Weaver architecture replaces the 90° phase shift network with quadrature mixing. To formulate the circuit's behavior, we begin with $x_A(t)$ and $x_B(t)$ as given by Eqs. (4.70) and (4.71), respectively, and perform the second

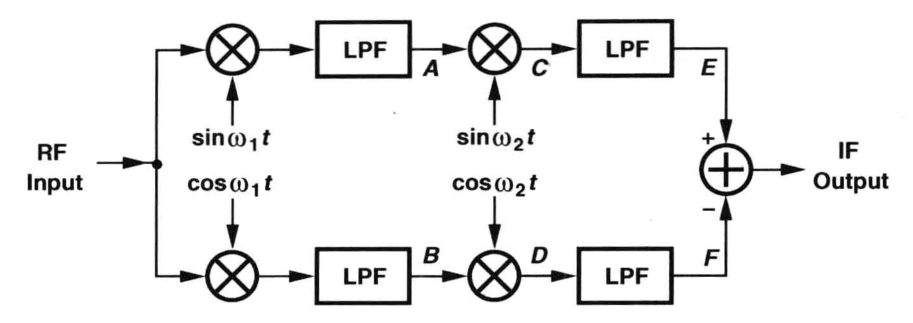


Figure 4.67 *Weaver architecture.*

quadrature mixing operation, arriving at

$$x_C(t) = \frac{A_{sig}}{4}\cos[(\omega_c - \omega_1 - \omega_2)t + \phi_{sig}] + \frac{A_{im}}{4}\cos[(\omega_{im} - \omega_1 - \omega_2)t + \phi_{im}] \\ + \frac{A_{sig}}{4}\cos[(\omega_c - \omega_1 + \omega_2)t + \phi_{sig}] + \frac{A_{im}}{4}\cos[(\omega_{im} - \omega_1 + \omega_2)t + \phi_{im}] \quad (4.84)$$

$$x_D(t) = -\frac{A_{sig}}{4}\cos[(\omega_c - \omega_1 - \omega_2)t + \phi_{sig}] - \frac{A_{im}}{4}\cos[(\omega_{im} - \omega_1 - \omega_2)t + \phi_{im}] \\ + \frac{A_{sig}}{4}\cos[(\omega_c - \omega_1 + \omega_2)t + \phi_{sig}] + \frac{A_{im}}{4}\cos[(\omega_{im} - \omega_1 + \omega_2)t + \phi_{im}]. \quad (4.85)$$

Should these results be added or subtracted? Let us assume low-side injection for both mixing stages. Thus, $\omega_{im} < \omega_1$ and $\omega_1 - \omega_{im} > \omega_2$ (Fig. 4.68). Also, $\omega_1 - \omega_{im} + \omega_2 > \omega_1 - \omega_{im} - \omega_2$. The low-pass filters following points C and D in Fig. 4.67 must therefore remove the components at $\omega_1 - \omega_{im} + \omega_2$ $(= \omega_c - \omega_1 + \omega_2)$, leaving only those at $\omega_1 - \omega_{im} - \omega_2$ $(= \omega_c - \omega_1 - \omega_2)$. That is, the *second* and *third* terms in Eqs. (4.84) and (4.85) are filtered. Upon subtracting $x_F(t)$ from $x_E(t)$, we obtain

$$x_E(t) - x_F(t) = \frac{A_{sig}}{2}\cos[(\omega_c - \omega_1 - \omega_2)t + \phi_{sig}]. \quad (4.86)$$

The image is therefore removed. In Problem 4.18, we consider the other three combinations of low-side and high-side injection so as to determine whether the outputs must be added or subtracted.

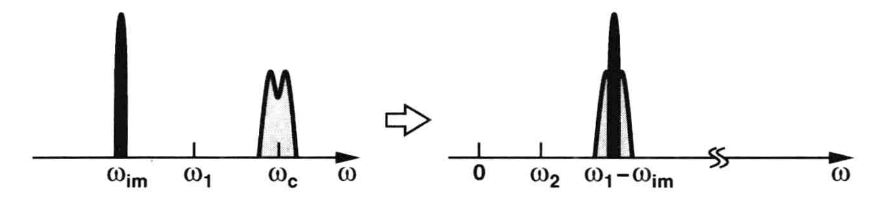


Figure 4.68 *RF and IF spectra in Weaver architecture.*

Example 4.32

Perform the above analysis graphically. Assume low-side injection for both mixing stages.

Solution:

Recall from Fig. 4.60(b) that low-side injection mixing with a sine multiplies the spectrum by $+(j/2)\text{sgn}(\omega)$. Beginning with the spectra of Fig. 4.61 and mixing them with $\sin\omega_2 t$ and $\cos\omega_2 t$, we arrive at the spectra shown in Fig. 4.69. Subtraction of $X_F(f)$ from $X_E(f)$ thus yields the signal and removes the image.

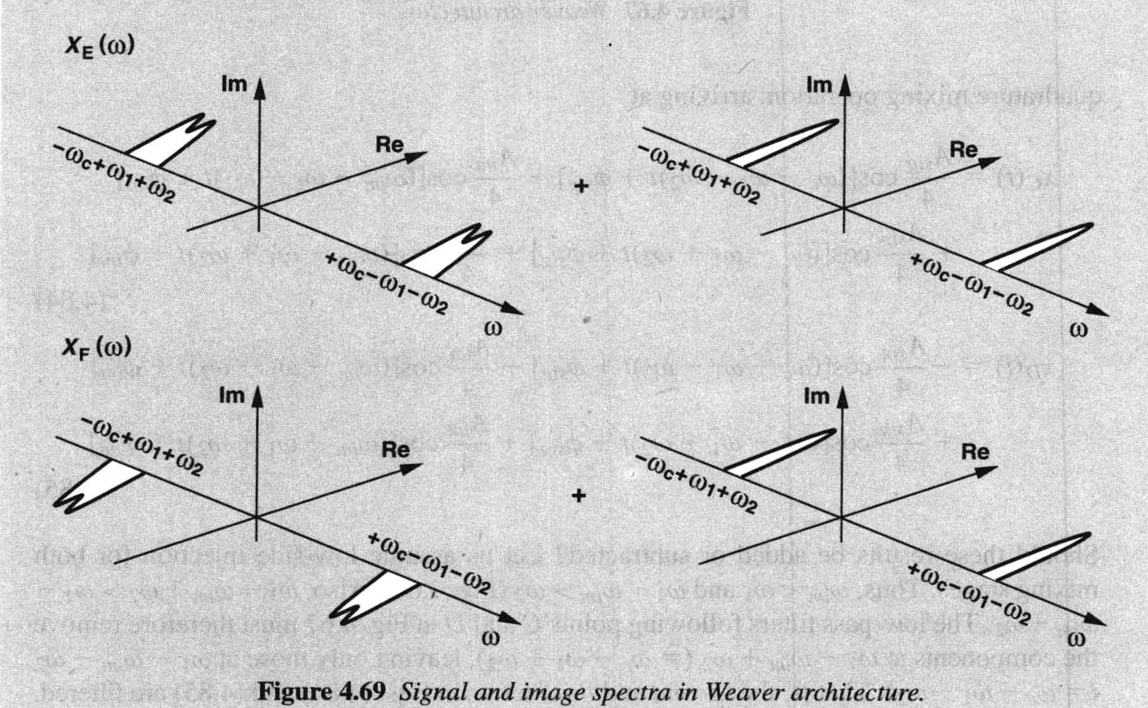


Figure 4.69 *Signal and image spectra in Weaver architecture.*

While employing two more mixers and one more LO than the Hartley architecture, the Weaver topology avoids the issues related to *RC-CR* networks: resistance and capacitance variations, degradation of IRR as the frequency departs from $1/(R_1C_1)$, attenuation, and noise. Also, if the IF mixers are realized in active form (Chapter 6), their outputs are available in the current domain and can be summed directly. Nonetheless, the IRR is still limited by mismatches, typically falling below 40 dB.

The Weaver architecture must deal with a *secondary* image if the second IF is not zero. Illustrated in Fig. 4.70, this effect arises if a component at $2\omega_2 - \omega_{in} + 2\omega_1$ accompanies the RF signal. Downconversion to the first IF translates this component to $2\omega_2 - \omega_{in} + \omega_1$, i.e., image of the signal with respect to ω_2, and mixing with ω_2 brings it to $\omega_2 - \omega_{in} + \omega_1$, the same IF at which the signal appears. For this reason, the second downconversion preferably

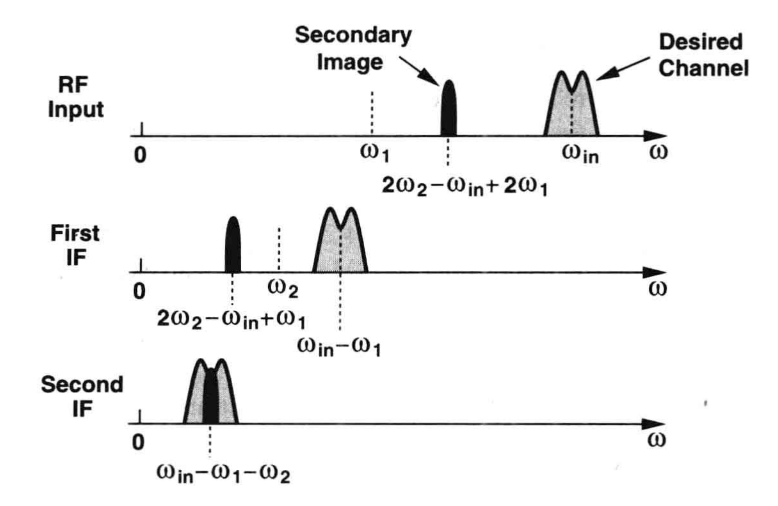


Figure 4.70 *Secondary image in Weaver architecture.*

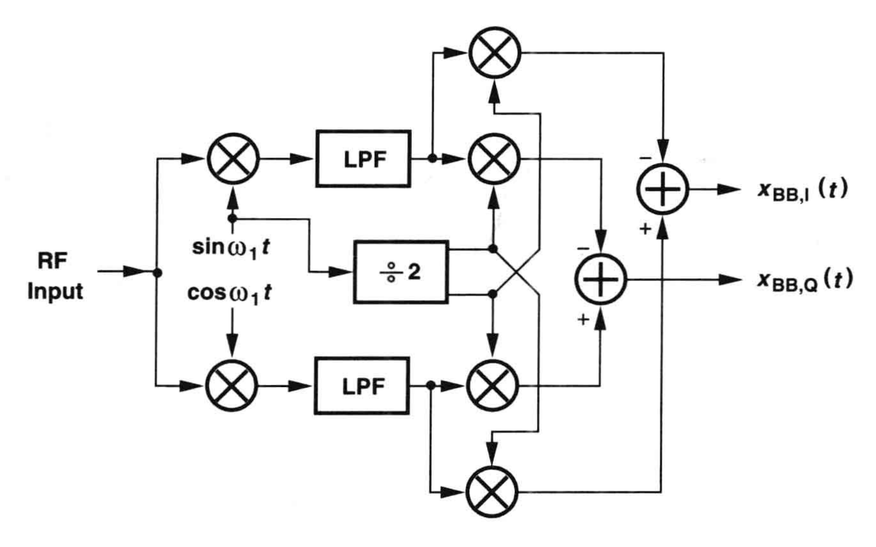


Figure 4.71 *Double quadrature downconversion Weaver architecture to produce baseband outputs.*

produces a zero IF, in which case it must perform quadrature separation as well. Figure 4.71 shows an example [6], where the second LO is derived from the first by frequency division.

The Weaver topology also suffers from mixing spurs in both downconversion steps. In particular, the harmonics of the second LO frequency may downconvert interferers from the first IF to baseband.

Calibration For image-rejection ratios well above 40 dB, the Hartley or Weaver architectures must incorporate calibration, i.e., a method of cancelling the gain and phase mismatches. A number of calibration techniques have been reported [7, 9].

4.2.5 Low-IF Receivers

In our study of heterodyne receivers, we noted that it is undesirable to place the image within the signal band because the image thermal noise of the antenna, the LNA, and the input stage of the RF mixer would raise the overall noise figure by approximately 3 dB.[22] In "low-IF receivers," the image indeed falls in the band but is suppressed by image rejection operations similar to those described in Section 4.2.4. To understand the motivation for the use of low-IF architectures, let us consider a GSM receiver as an example. As explained in Section 4.2.3, direct conversion of the 200-kHz desired channel to a zero IF may significantly corrupt the signal by $1/f$ noise. Furthermore, the removal of the dc offset by means of a high-pass filter proves difficult. Now suppose the LO frequency is placed at the *edge* of the desired (200-kHz) channel [Fig. 4.72(a)], thereby translating the RF signal to an IF of 100 kHz. With such an IF, and because the signal carries little information near the edge, the $1/f$ noise penalty is much less severe. Also, on-chip high-pass filtering of the signal becomes feasible. Called a "low-IF receiver," this type of system is particularly attractive for narrow-channel standards.

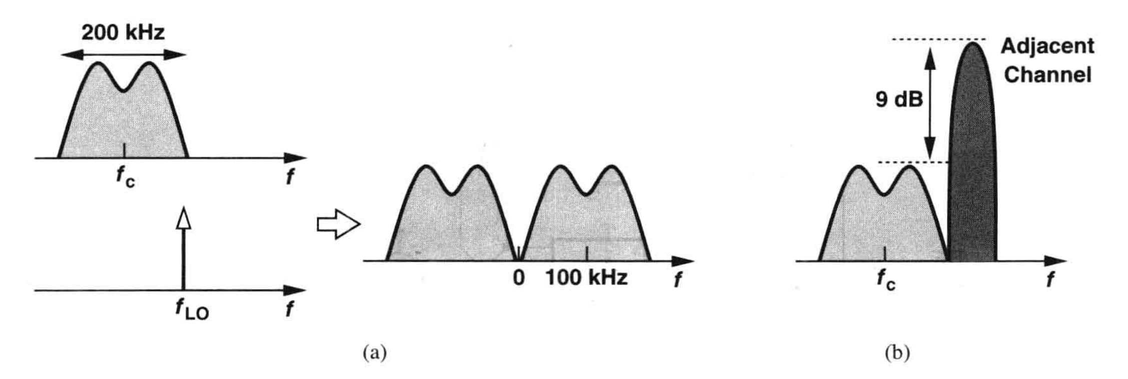


Figure 4.72 *(a) Spectra in a low-IF receiver, (b) adjacent-channel specification in GSM.*

The heterodyne downconversion nonetheless raises the issue of the image, which in this case falls in the adjacent channel. Fortunately, the GSM standard requires that receivers tolerate an adjacent channel only 9 dB above the desired channel (Chapter 3) [Fig. 4.72(b)]. Thus, an image-reject receiver with a moderate IRR can lower the image to well below the signal level. For example, if IRR = 30 dB, the image remains 21 dB below the signal.

Example 4.33

Repeat Example 4.24 for a low-IF receiver.

Solution:

Assuming that high-pass filtering of dc offsets also removes the flicker noise up to roughly 20 kHz, we integrate the noise from 20 kHz to 200 kHz (Fig. 4.73):

22. This occurs because the entire signal band must see a flat frequency response in the antenna/LNA/mixer chain.

Example 4.33 (*Continued*)

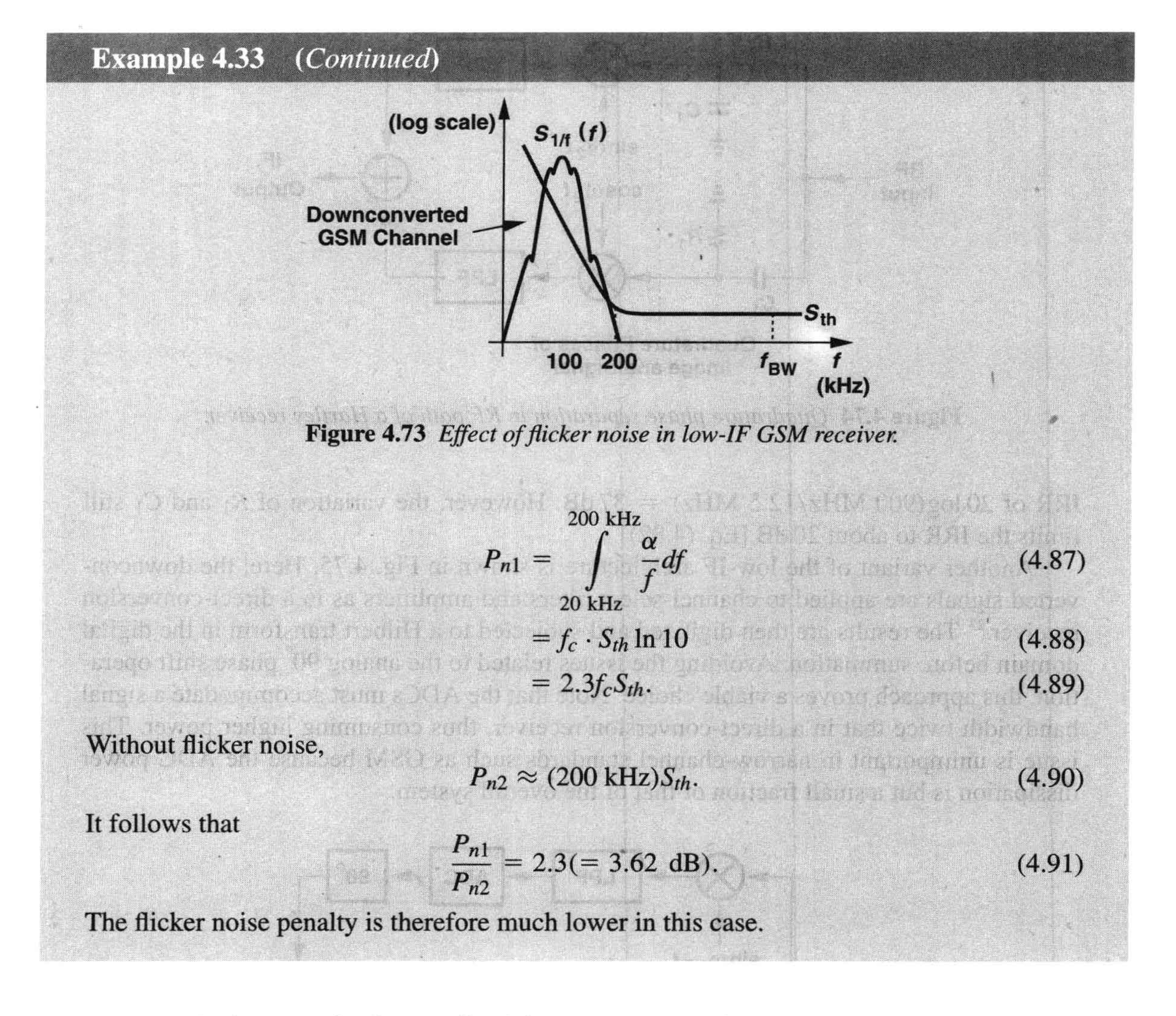


Figure 4.73 *Effect of flicker noise in low-IF GSM receiver.*

$$P_{n1} = \int_{20\ \text{kHz}}^{200\ \text{kHz}} \frac{\alpha}{f} df \tag{4.87}$$

$$= f_c \cdot S_{th} \ln 10 \tag{4.88}$$

$$= 2.3 f_c S_{th}. \tag{4.89}$$

Without flicker noise,

$$P_{n2} \approx (200\ \text{kHz}) S_{th}. \tag{4.90}$$

It follows that

$$\frac{P_{n1}}{P_{n2}} = 2.3 (= 3.62\ \text{dB}). \tag{4.91}$$

The flicker noise penalty is therefore much lower in this case.

How is image rejection realized in a low-IF receiver? The Hartley architecture employing the *RC-CR* network (Fig. 4.64) appears to be a candidate, but the IF spectrum in a low-IF RX may extend to zero frequency, making it impossible to maintain a high IRR across the signal bandwidth. (The high-pass Section exhibits zero gain near frequency!) While avoiding this issue, the Weaver architecture must deal with the secondary image if the second IF is not zero or with flicker noise if it is.

One possible remedy is to move the 90° phase shift in the Hartley architecture from the IF path to the RF path. Illustrated in Fig. 4.74, the idea is to first create the quadrature phases of the RF signal and the image and subsequently perform another Hilbert transform by means of quadrature mixing. We also recognize some similarity between this topology and the Weaver architecture: both multiply quadrature components of the signal and the image by the quadrature phases of the LO and sum the results, possibly in the current domain. Here, the *RC-CR* network is centered at a high frequency and can maintain a reasonable IRR across the band. For example, for the 25-MHz receive band of 900-MHz GSM, if $(2\pi R_1 C_1)^{-1}$ is chosen equal to the center frequency, then Eq. (4.83) implies an

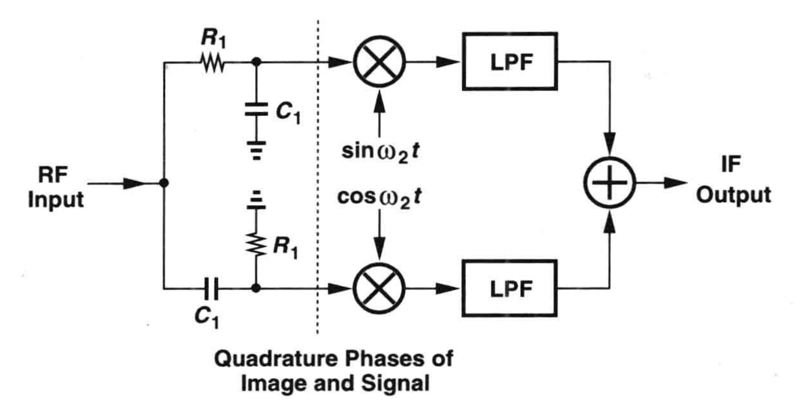


Figure 4.74 *Quadrature phase separation in RF path of a Hartley receiver.*

IRR of $20\log(900\text{ MHz}/12.5\text{ MHz}) = 37\text{ dB}$. However, the variation of R_1 and C_1 still limits the IRR to about 20 dB [Eq. (4.82)].

Another variant of the low-IF architecture is shown in Fig. 4.75. Here, the downconverted signals are applied to channel-select filters and amplifiers as in a direct-conversion receiver.[23] The results are then digitized and subjected to a Hilbert transform in the digital domain before summation. Avoiding the issues related to the analog 90° phase shift operation, this approach proves a viable choice. Note that the ADCs must accommodate a signal bandwidth twice that in a direct-conversion receiver, thus consuming higher power. This issue is unimportant in narrow-channel standards such as GSM because the ADC power dissipation is but a small fraction of that of the overall system.

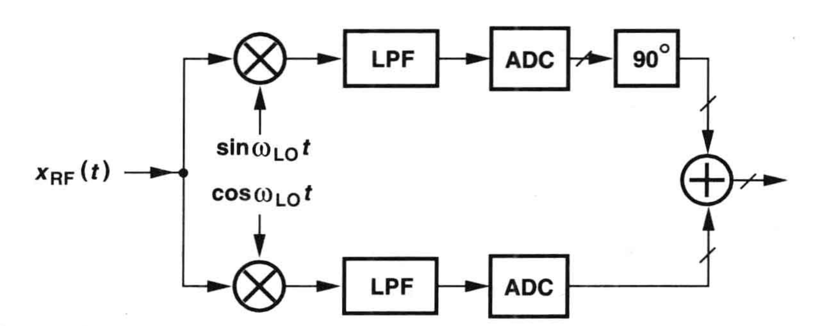


Figure 4.75 *Low-IF receiver with 90° phase shift in digital domain.*

Let us summarize our thoughts thus far. If a low-IF receiver places the image in the adjacent channel, then it cannot employ an *RC-CR* 90° phase shift after downconversion. Also, a 90° circuit in the RF path still suffers from *RC* variation. For these reasons, the concept of "low IF" can be extended to any downconversion that places the image within

23. The channel-select filters must, however, provide a bandwidth equal to the RF signal bandwidth rather than half of it [Fig. 4.72(a)].

the band so that the IF is significantly *higher* than the signal bandwidth, possibly allowing the use of an *RC-CR* network—but not so high as to unduly burden the ADC. Of course, since the image no longer lies in the adjacent channel, a substantially higher IRR may be required. Some research has therefore been expended on low-IF receivers with high image rejection. Such receivers often employ "polyphase filters" (PPFs) [10, 11].

Polyphase Filters Recall from Section 4.2.4 that heterodyne quadrature downconversion subjects the signal to low-side injection and the image to high-side injection, or vice versa, thus creating the Hilbert transform of one and the negative Hilbert transform of the other. Now let us consider the circuit shown in Fig. 4.76(a), where V_{out} can be viewed as a weighted sum of V_1 and V_2:

$$V_{out} = \frac{V_1 + R_1 C_1 s V_2}{R_1 C_1 s + 1}. \tag{4.92}$$

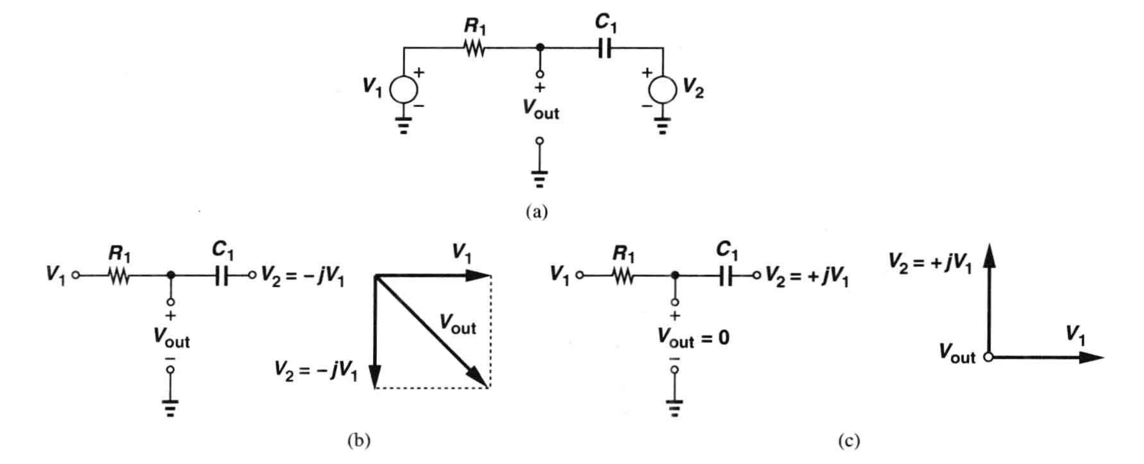


Figure 4.76 *(a) Simple RC circuit, (b) output in response to V_1 and $-jV_1$, (c) output in response to V_1 and $+jV_1$.*

We consider two special cases.

1. The voltage V_2 is the Hilbert transform of V_1; in phasor form, $V_2 = -jV_1$ (Example 4.26). Consequently, for $s = j\omega$,

$$V_{out} = V_1 \frac{R_1 C_1 \omega + 1}{jR_1 C_1 \omega + 1}. \tag{4.93}$$

If $\omega = (R_1 C_1)^{-1}$, then $V_{out} = 2V_1/(1 + j) = V_1(1 - j)$. That is, $|V_{out}| = \sqrt{2}V_1$ and $\angle V_{out} = \angle V_{in} - 45^\circ$ [Fig. 4.76(b)]. In this case, the circuit simply computes the vector summation of V_1 and $V_2 = -jV_1$. We say the circuit rotates by -45° the voltage sensed by the resistor.

2. The voltage V_2 is the *negative* Hilbert transform of V_1, i.e., $V_2 = +jV_1$. For $s = j\omega$,

$$V_{out} = V_1 \frac{-R_1 C_1 \omega + 1}{jR_1 C_1 \omega + 1}. \tag{4.94}$$

Interestingly, if $\omega = (R_1C_1)^{-1}$, then $V_{out} = 0$ [Fig. 4.76(c)]. Intuitively, we can say that C_1 rotates V_2 by another 90° so that the result cancels the effect of V_1 at the output node. The reader is encouraged to arrive at these conclusions using superposition.

In summary, the series branch of Fig. 4.76(a) rotates V_1 by -45° (to produce V_{out}) if $V_2 = -jV_1$ and *rejects* V_1 if $V_2 = +jV_1$. The circuit can therefore distinguish between the signal and the image if it follows a quadrature downconverter.

Example 4.34

Extend the topology of Fig. 4.76(a) if V_1 and $-jV_1$ are available in differential form and construct an image-reject receiver.

Solution:

Figure 4.77(a) shows the arrangement and the resulting phasors if $R_1 = R_2 = R$ and $C_1 = C_2 = C$. The connections to quadrature downconversion mixers are depicted in Fig. 4.77(b).

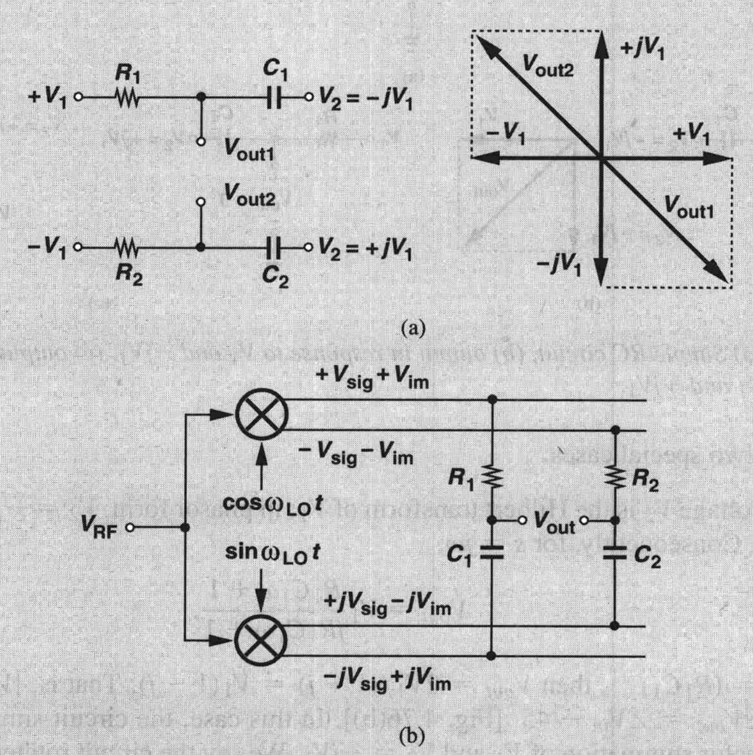


Figure 4.77 *(a) RC circuit sensing differential inputs, (b) quadrature downconverter driving RC network of part (a).*

In contrast to the Hartley architecture of Fig. 4.62, the circuit of Fig. 4.77(b) avoids an explicit voltage adder at the output. Nonetheless, this arrangement still suffers from RC variations and a narrow bandwidth. In fact, at an IF of $\omega = (R_1C_1)^{-1} + \Delta\omega$, Eq. (4.94) yields the residual image as

$$|V_{out}| \approx |V_1| \frac{RC\Delta\omega}{\sqrt{2 + 2RC\Delta\omega}} \tag{4.95}$$

$$\approx |V_1| \frac{RC\Delta\omega}{\sqrt{2}}, \tag{4.96}$$

where it is assumed that $\Delta\omega \ll \omega$.

In the next step of our development of polyphase filters, let us now redraw the circuit of Fig. 4.77(a) and add two branches to it as shown in Fig. 4.78(a). Here, the capacitors are equal and so are the resistors. The top and bottom branches still produce differential outputs, but how about the left and right branches? Since R_3 and C_3 compute the weighted sum of $+jV_1$ and $+V_1$, we observe from Fig. 4.76(b) that V_{out3} is 45° more *negative* than $+jV_1$. By the same token, V_{out4} is 45° more negative than $-jV_1$. Figure 4.78(b) depicts the resulting phasors at $\omega = (R_1C_1)^{-1}$, suggesting that the circuit produces quadrature outputs that are 45° out of phase with respect to the quadrature inputs.

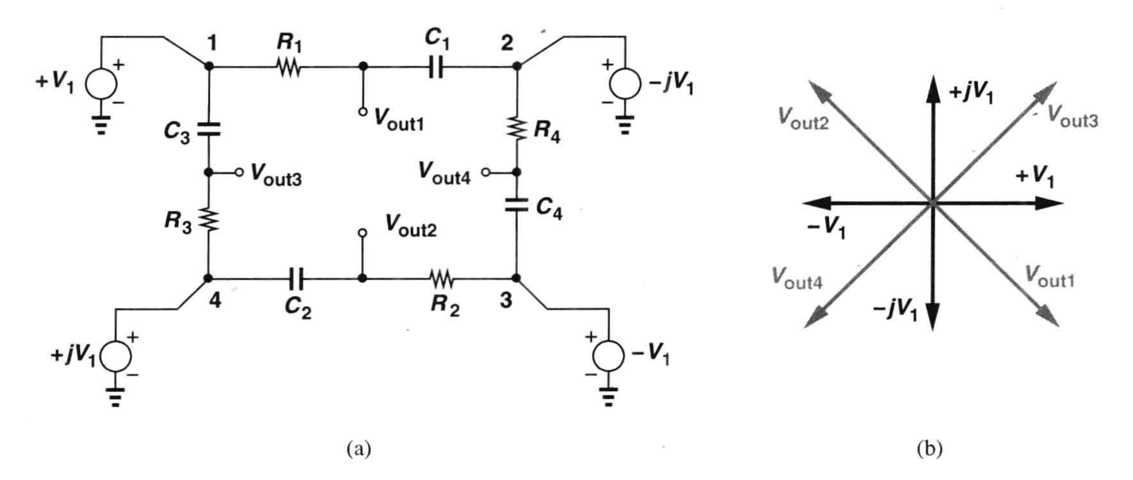


Figure 4.78 *(a) RC network sensing differential quadrature phases, (b) resulting outputs.*

Example 4.35

The outputs of a quadrature downconverter contain the signal, V_{sig}, and the image, V_{im}, and drive the circuit of Fig. 4.78(a) as shown in Fig. 4.79(a). Determine the outputs, assuming all capacitors are equal to C and all resistors equal to R.

(*Continues*)

Example 4.35 *(Continued)*

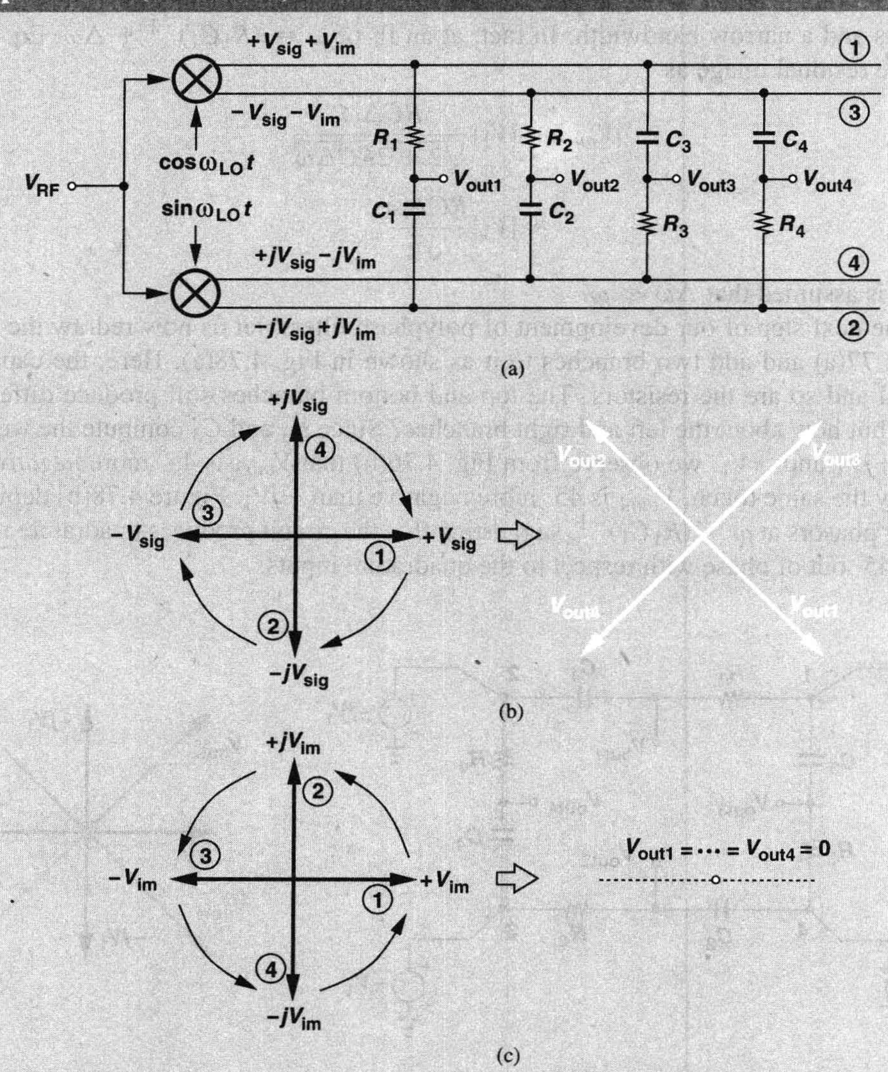


Figure 4.79 *(a) Quadrature downconverter driving RC sections, (b) resulting signal output, (c) resulting image output.*

Solution:

The quadrature downconverter produces $+V_{sig} + V_{im}$, $-V_{sig} - V_{im}$, $+jV_{sig} - jV_{im}$, and $-jV_{sig} + jV_{im}$. At $\omega = (RC)^{-1}$, the branch R_1C_1 rotates $+V_{sig}$ by -45° to generate V_{out1}. Similarly, R_2C_2 rotates $-V_{sig}$ by 45° to generate V_{out2}, etc. [Fig. 4.79(b)]. The image components, on the other hand, yield a zero output [Fig. 4.79(c)]. The key point here is that, if we consider the sequence of nodes 1, 2, 3, and 4, we observe that V_{sig} rotates *clockwise* by

Example 4.35 *(Continued)*

90° from one node to the next, whereas V_{im} rotates *counterclockwise* by 90° from one node to the next. The circuit therefore exhibits an asymmetry in its response to the "sequence" of the four inputs.

The multiphase circuit of Fig. 4.78(a) is called a "sequence-asymmetric polyphase filter" [8]. Since the signal and the image arrive at the inputs with different sequences, one is passed to the outputs while the other is suppressed. But what happens if $\omega \neq (RC)^{-1}$? Substituting $\omega = (R_1C_1)^{-1} + \Delta\omega$ in Eq. (4.93), we have

$$V_{out1} = V_{sig}\frac{2 + RC\Delta\omega}{1 + j(1 + RC\Delta\omega)}, \tag{4.97}$$

and hence,

$$|V_{out1}|^2 = |V_{sig}|^2\frac{4 + 4RC\Delta\omega + R^2C^2\Delta\omega^2}{2 + 2RC\Delta\omega + R^2C^2\Delta\omega^2} \tag{4.98}$$

$$\approx 2|V_{sig}|^2\left(1 + RC\Delta\omega + \frac{R^2C^2\Delta\omega^2}{4}\right)\left(1 - RC\Delta\omega - \frac{R^2C^2\Delta\omega^2}{2}\right) \tag{4.99}$$

$$\approx 2|V_{sig}|^2\left(1 - \frac{5}{4}R^2C^2\Delta\omega^2\right). \tag{4.100}$$

That is,

$$|V_{out1}| \approx \sqrt{2}|V_{sig}|\left(1 - \frac{5}{8}R^2C^2\Delta\omega^2\right). \tag{4.101}$$

The phase of V_{out1} is obtained from (4.97) as

$$\angle V_{out1} = \angle V_{sig} - \tan^{-1}(1 + RC\Delta\omega). \tag{4.102}$$

Since $\tan^{-1}(1 + RC\Delta\omega) \approx \pi/4 + RC\Delta\omega/2$ for $RC\Delta\omega \ll 1$ rad,

$$\angle V_{out1} = \angle V_{sig} - \left(\frac{\pi}{4} + \frac{RC\Delta\omega}{2}\right). \tag{4.103}$$

Figure 4.80(a) illustrates the effect on all four phases of the signal, implying that the outputs retain their differential and quadrature relationship.

For the image, we return to Eq. (4.96) and note that the four outputs have a magnitude equal to $V_{im}RC\Delta\omega/\sqrt{2}$ and phases similar to those of the signal components in Fig. 4.80(a). The output image phasors thus appear as shown in Fig. 4.80(b). The reader is encouraged to prove that V_{out1} is at $-45^{\circ} - RC\Delta\omega/2$ and V_{out3} at $-135^{\circ} - RC\Delta\omega/2$.

An interesting observation in Fig. 4.80 is that the output signal and image components exhibit *opposite* sequences [10, 11]. We therefore expect that if this polyphase filter is

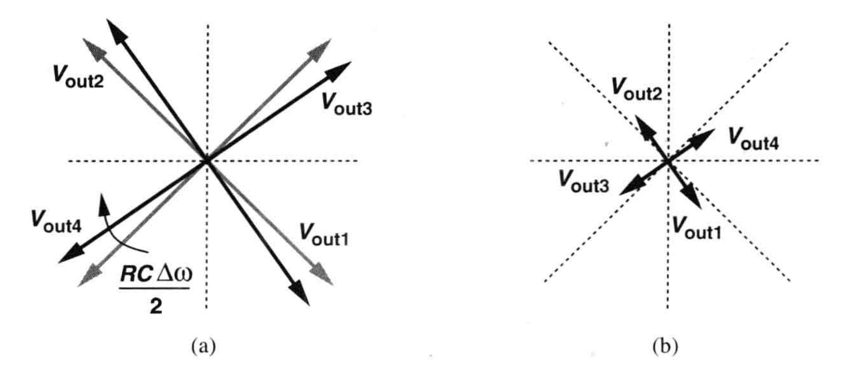


Figure 4.80 *Effect of polyphase filter at a frequency offset of $\Delta\omega$ for (a) signal, and (b) image.*

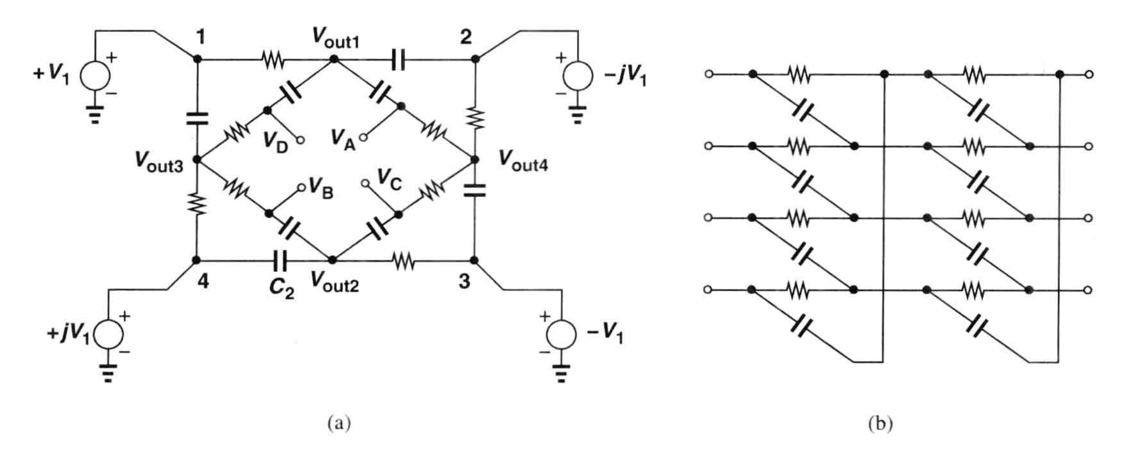


Figure 4.81 *(a) Cascaded polyphase sections, (b) alternative drawing.*

followed by another, then the image can be further suppressed. Figure 4.81(a) depicts such a cascade and Fig. 4.81(b) shows an alternative drawing that more easily lends itself to cascading.

We must now answer two questions: (1) how should we account for the loading of the second stage on the first? (2) how are the *RC* values chosen in the two stages? To answer the first question, consider the equivalent circuit shown in Fig. 4.82, where Z_1-Z_4 represent the *RC* branches in the second stage. Intuitively, we note that Z_1 attempts to "pull" the phasors V_{out1} and V_{out3} toward each other, Z_2 attempts to pull V_{out1} and V_{out4} toward each other, etc. Thus, if $Z_1 = \cdots = Z_4 = Z$, then, V_{out1}–V_{out4} experience no rotation, but the loading may reduce their *magnitudes*. Since the angles of V_{out1}–V_{out4} remain unchanged, we can express them as $\pm\alpha(1 \pm j)V_1$, where α denotes the attenuation due to the loading of the second stage. The currents drawn from node X by Z_1 and Z_2 are therefore equal to $[\alpha(1 - j)V_1 - \alpha(1 + j)V_1]/Z_1$ and $[\alpha(1 - j)V_1 + \alpha(1 + j)V_1]/Z_2$, respectively. Summing

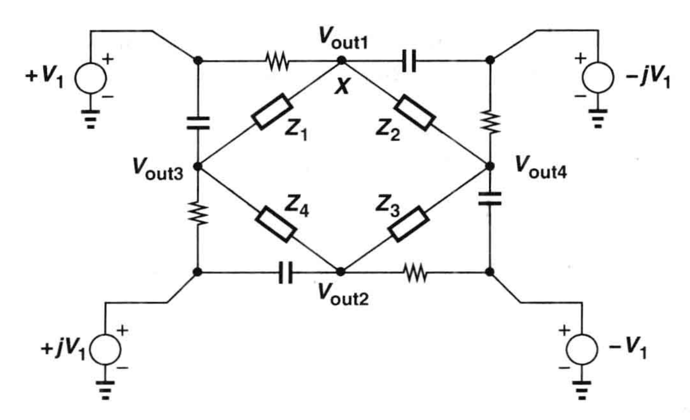


Figure 4.82 *Effect of loading of second polyphase section.*

all of the currents flowing out of node X and equating the result to zero, we have

$$\frac{\alpha(1-j)V_1 - V_1}{R} + [\alpha(1-j)V_1 + jV_1]Cj\omega + \frac{\alpha(1-j)V_1 - \alpha(1+j)V_1}{Z} + \frac{\alpha(1-j)V_1 + \alpha(1+j)V_1}{Z} = 0. \tag{4.104}$$

This equality must hold for any nonzero value of V_1. If $RC\omega = 1$, the expression reduces to

$$2\alpha - 2 + \frac{2\alpha(1-j)R}{Z} = 0. \tag{4.105}$$

That is,

$$\alpha = \frac{Z}{Z + (1-j)R}. \tag{4.106}$$

For example, if $Z = R + (jC\omega)^{-1}$, then $\alpha = 1/2$, revealing that loading by an identical stage attenuates the outputs of the first stage by a factor of 2.

Example 4.36

If $Z = R + (jC\omega)^{-1}$ and $RC\omega = 1$, determine V_A in Fig. 4.81(a).

Solution:

We have $V_{out1} = (1/2)(1 - jV_1)$ and $V_{out4} = (1/2)(-1-j)V_1$, observing that V_{out1} and V_{out2} have the same phase relationship as V_1 and V_2 in Fig. 4.76(b). Thus, V_A is simply the vector sum of V_{out1} and V_{out4}:

$$V_A = -jV_1. \tag{4.107}$$

In comparison with Fig. 4.76(b), we note that a two-section polyphase filter produces an output whose magnitude is $\sqrt{2}$ times smaller than that of a single-section counterpart. We say each section attenuates the signal by a factor of $\sqrt{2}$.

The second question relates to the choice of RC values. Suppose both stages employ $RC = R_0C_0$. Then, the cascade of two stages yields an image attenuation equal to the *square* of Eq. (4.95) at a frequency of $(R_0C_0)^{-1} + \Delta\omega$:

$$\left|\frac{V_{im,out}}{V_{im,in}}\right| \approx \frac{(R_0C_0\Delta\omega)^2}{2 + 2R_0C_0\Delta\omega}, \tag{4.108}$$

which reduces to $(R_0C_0\Delta\omega)^2/2$ for $\Delta\omega \ll (R_0C_0)^{-1}$. Figure 4.83 plots this behavior, comparing it with that of a single section.

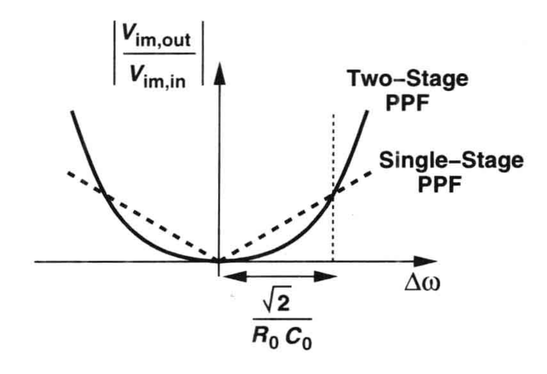


Figure 4.83 *Image rejection for single-stage and two-stage polyphase filters.*

What happens if the two stages use different time constants? In particular, for a given IF of ω_0, let us assume that the time constants in the first and second stages are respectively equal to R_1C_1 and R_2C_2 such that $\omega_0 - (R_1C_1)^{-1} = (R_2C_2)^{-1} - \omega_0$, i.e., the center frequencies are shifted up and down. From Eq. (4.96), we plot the image rejections of the two stages as shown in Fig. 4.84(a).[24] The product of these two functions is a parabola crossing zero at $\omega_1 = (R_1C_1)^{-1}$ and $\omega_2 = (R_2C_2)^{-1}$ [Fig. 4.84(b)]. The reader can prove that the attenuation at ω_0 is equal to $(\omega_1 - \omega_2)^2/(8\omega_1\omega_2)$, which must be chosen sufficiently small. The reader can also show that for an attenuation of 60 dB, $\omega_1 - \omega_2$ cannot exceed approximately 18% of ω_0.

The advantage of splitting the cut-off frequencies of the two stages is the wider achievable bandwidth. Figure 4.85 plots the image rejection for $\omega_1 = \omega_2 = \omega_0$ and $\omega_1 \neq \omega_2$.

The cascading of polyphase filter sections entails both attenuation and additional thermal noise. To alleviate the former, the resistors in the latter stages can be chosen larger than those in the former, but at the cost of higher noise. For this reason, polyphase filters are only occasionally used in RF receivers. In low-IF architectures, the polyphase filters can be realized as "complex filters" so as to perform channel-selection filtering [12].

Double-Quadrature Downconversion In our study of the Hartley architecture, we noted that mismatches arise in both the RF signal path and the LO path. A method of reducing the effect of mismatches incorporates "double-quadrature" downconversion [10]. Illustrated in Fig. 4.86, the circuit decomposes the RF signal into quadrature components, performs

24. For clarity, the plots are allowed to be negative even though Eq. (4.96) contains absolute values.

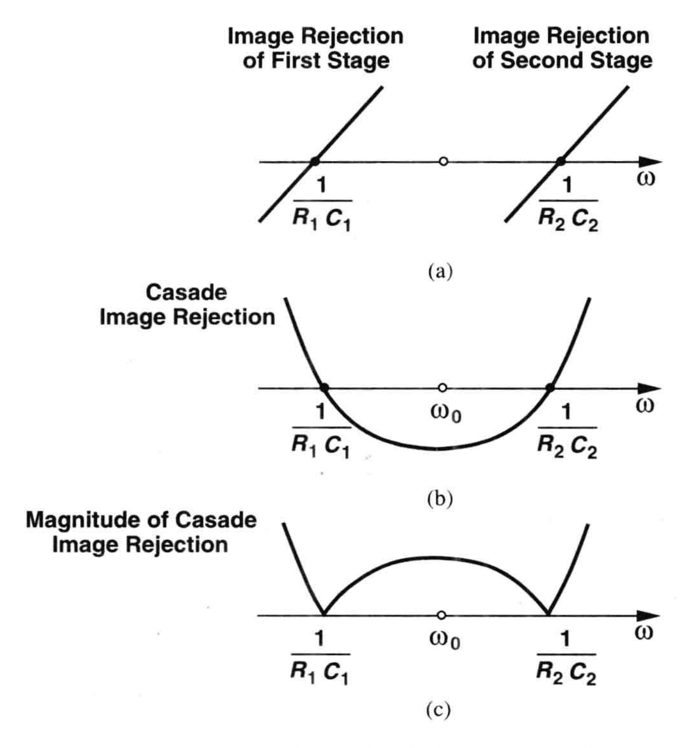


Figure 4.84 *(a) Image rejections of two unidentical polyphase stages, (b) cascade image rejection, (c) magnitude of cascade image rejection.*

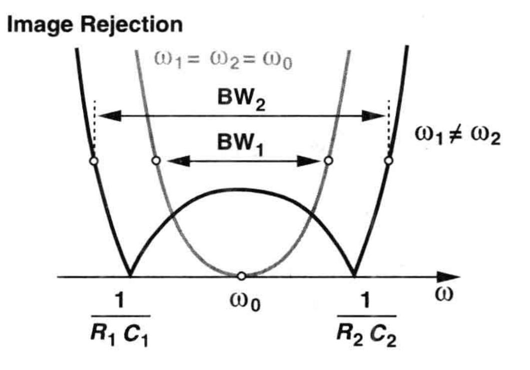


Figure 4.85 *Comparison of image rejections of two identical and unidentical polyphase stages.*

quadrature downconversion on *each* of the RF components, and subtracts and adds the results to produce net quadrature IF outputs. It can be shown [10] that the overall gain and phase mismatches of this topology are given by

$$\frac{\Delta A}{A} = \frac{\Delta A_{RF}}{A_{RF}} \cdot \frac{\Delta A_{LO}}{A_{LO}} + \frac{\Delta G_{mix}}{G_{mix}} \tag{4.109}$$

$$\tan(\Delta\phi) = \tan(\Delta\phi_{RF}) \cdot \tan(\Delta\phi_{LO}) + \frac{\tan(\Delta\phi_{mix})}{2}, \tag{4.110}$$

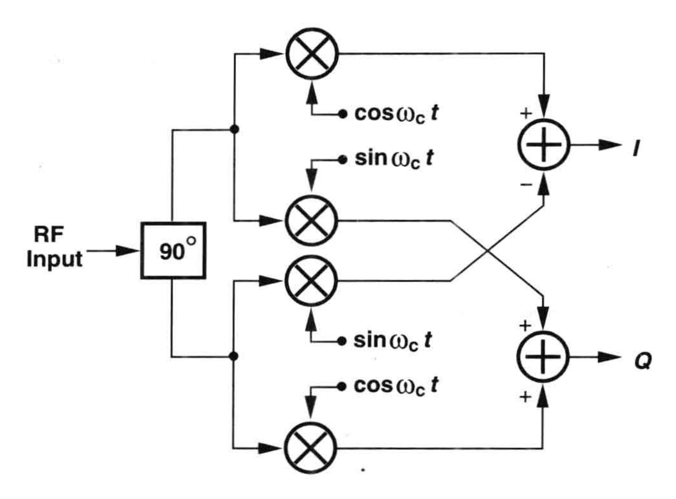


Figure 4.86 *Low-IF receiver with double quadrature downconverter.*

where $\Delta A_{RF}/A_{RF}$ and $\Delta A_{LO}/A_{LO}$ denote amplitude mismatches in the RF and LO paths, respectively, and $\Delta\phi_{RF}$ and $\Delta\phi_{LO}$ are the corresponding phase mismatches. The quantities $\Delta G_{mix}/G_{mix}$ and $\Delta\phi_{mix}$ denote the conversion gain mismatch and phase mismatch of the mixers, respectively. Thus, the IRR is only limited by mixer mismatches because the first terms on the right-hand sides of (4.109) and (4.110) are very small.

Equations (4.109) and (4.110) reveal that the IRR of the double-quadrature architecture is likely to fall well below 60 dB. This is because, even with no phase mismatch, $\Delta G_{mix}/G_{mix}$ must remain less than 0.1% for IRR = 60 dB, a very stringent requirement even for matching of simple resistors. Calibration of the mismatches can therefore raise the IRR more robustly [7, 9].

4.3 TRANSMITTER ARCHITECTURES

4.3.1 General Considerations

An RF transmitter performs modulation, upconversion, and power amplification. In most of today's systems, the data to be transmitted is provided in the form of quadrature baseband signals. For example, from Chapter 3 the GMSK waveform in GSM can be expanded as

$$x_{GMSK}(t) = A\cos[\omega_c t + m\int x_{BB}(t) * h(t)dt] \tag{4.111}$$

$$= A\cos\omega_c t\cos\phi - A\sin\omega_c t\sin\phi, \tag{4.112}$$

where

$$\phi = m\int x_{BB} * h(t)dt. \tag{4.113}$$

Thus, $\cos\phi$ and $\sin\phi$ are produced from $x_{BB}(t)$ by the digital baseband processor, converted to analog form by D/A converters, and applied to the transmitter.

We have seen the need for baseband pulse shaping in Chapter 3 and in the above equations for GMSK: each rectangular data pulse must be transformed to a smoother pulse.

Since pulse shaping in the analog domain, especially at low frequencies, requires bulky filters, each incoming pulse is mapped to the desired shape by a combination of digital and analog techniques. Illustrated in Fig. 4.87 is an example [13, 14], where the input pulse generates a sequence of addresses, e.g., it enables a counter, producing a set of levels from two read-only memories (ROMs). (We say the pulse is "oversampled.") These levels are subsequently converted to analog form, yielding the desired pulse shape at points A and B.

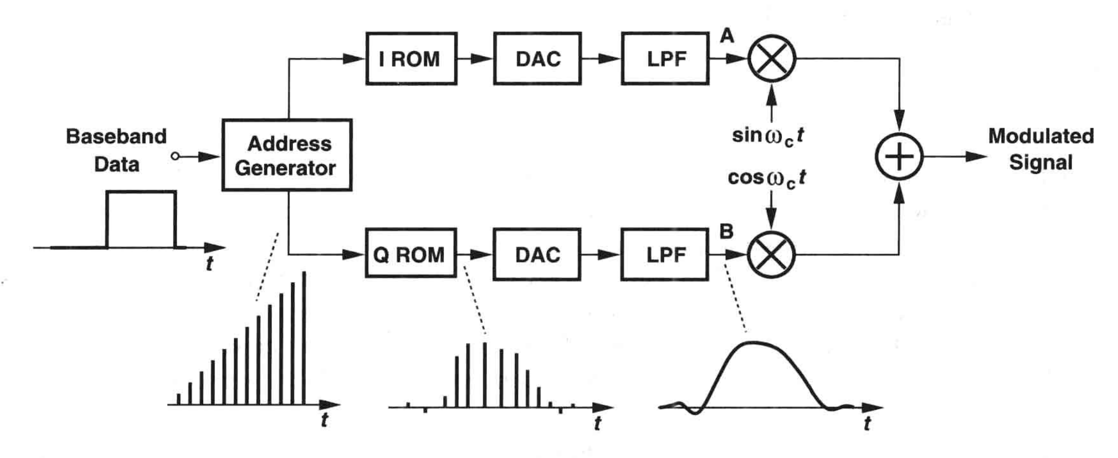


Figure 4.87 *Baseband pulse shaping.*

4.3.2 Direct-Conversion Transmitters

Most transmitter architectures are similar to the receiver topologies described in Section 4.2, but with the operations performed in "reverse" order. For example, an RX cascade of an LNA and quadrature downconversion mixers suggests a TX cascade of quadrature upconversion mixers and a PA.

The above expression of a GMSK waveform can be generalized to any narrowband modulated signal:

$$x(t) = A(t)\cos[\omega_c t + \phi(t)] \quad (4.114)$$

$$= A(t)\cos\omega_c t\cos[\phi(t)] - A(t)\sin\omega_c t\sin[\phi(t)]. \quad (4.115)$$

We therefore define the quadrature baseband signals as

$$x_{BB,I}(t) = A(t)\cos[\phi(t)] \quad (4.116)$$

$$x_{BB,Q}(t) = A(t)\sin[\phi(t)], \quad (4.117)$$

and construct the transmitter as shown in Fig. 4.88. Called a "direct-conversion" transmitter, this topology directly translates the baseband spectrum to the RF carrier by means of a "quadrature upconverter."[25] The upconverter is followed by a PA and a matching network, whose role is to provide maximum power delivery to the antenna and filter out-of-band

25. Also known as a "quadrature modulator" or a "vector modulator."

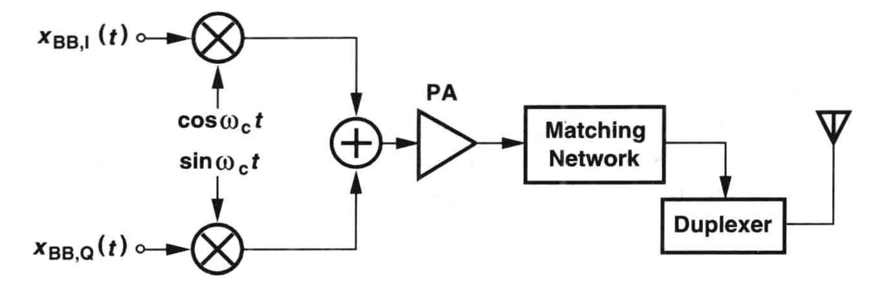


Figure 4.88 *Direct-conversion transmitter.*

components that result from the PA nonlinearity (Chapter 12). Note that since the baseband signal is produced in the transmitter and hence has a sufficiently large amplitude (several hundred millivolts), the noise of the mixers is much less critical here than in receivers.

The design of the TX begins with the PA. This circuit is crafted so as to deliver the required power to the antenna (and with adequate linearity if applicable). Employing large transistors to carry high currents, the PA exhibits a large input capacitance. Thus, a predriver is typically interposed between the upconverter and the PA to serve as a buffer.

Example 4.37

A student decides to omit the predriver and simply "scale up" the upconverter so that it can drive the PA directly. Explain the drawback of this approach.

Solution:

In order to scale up the upconverter, the width and bias current of each transistor are scaled up, and the resistor and inductor values are proportionally scaled down. For example, if the upconverter is modeled as a transconductance G_m and an output resistance R_{out} (Fig. 4.89),[26] then R_{out} can be reduced to yield adequate bandwidth with the input capacitance of the PA, and G_m can be enlarged to maintain a constant $G_m R_{out}$ (i.e., constant voltage swings). In practice, the upconverter employs a resonant LC load, but the same principles still apply.

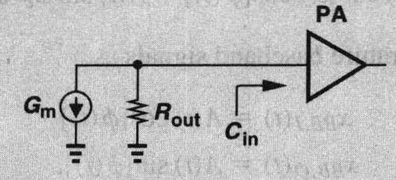


Figure 4.89 *Scaling of quadrature upconverter to drive the PA.*

The scaling of the transistors raises the capacitances seen at the baseband and LO ports of the mixers in Fig. 4.88. The principal issue here is that the LO now sees a large load capacitance, requiring its own buffers. Also, the two mixers consume a higher power.

26. In this conceptual model, we omit the frequency translation inherent in the upconverter.

Among various TX architectures studied in this chapter, the direct-conversion approach provides the most compact solution and a relatively "clean" output. That is, the output spectrum contains only the desired signal around the carrier frequency (and its harmonics) but no spurious components—an attribute similar to that of direct-conversion receivers. Nonetheless, direct upconversion entails other issues that demand attention.

I/Q Mismatch We noted in Section 4.2 that I/Q mismatch in direct-conversion receivers results in "cross-talk" between the quadrature baseband outputs or, equivalently, distortion in the constellation. We expect a similar effect in the TX counterpart. For example, as derived in Chapter 3, a phase mismatch of $\Delta\theta$ in the upconverter shifts the constellation points of a QPSK signal to $I = \pm\cos\Delta\theta$ and $Q = \pm 1 \pm \sin\Delta\theta$, just as observed for a direct-conversion receiver. The results obtained in Chapter 3 can be extended to include amplitude mismatch by writing

$$x(t) = \alpha_1(A_c + \Delta A_c)\cos(\omega_c t + \Delta\theta) + \alpha_2 A_c \sin\omega_c t \tag{4.118}$$

$$= \alpha_1(A_c + \Delta A_c)\cos\Delta\theta\cos\omega_c t + [\alpha_2 A_c - \alpha_1(A_c + \Delta A_c)\sin\Delta\theta]\sin\omega_c t. \tag{4.119}$$

Since α_1 and α_2 assume ± 1 values, the normalized coefficients of $\cos\omega_c t$ and $\sin\omega_c t$ appear as follows for the four points in the constellation:

$$\beta_1 = +\left(1 + \frac{\Delta A_c}{A_c}\right)\cos\Delta\theta,\ \beta_2 = 1 - \left(1 + \frac{\Delta A_c}{A_c}\right)\sin\Delta\theta \tag{4.120}$$

$$\beta_1 = +\left(1 + \frac{\Delta A_c}{A_c}\right)\cos\Delta\theta,\ \beta_2 = -1 - \left(1 + \frac{\Delta A_c}{A_c}\right)\sin\Delta\theta \tag{4.121}$$

$$\beta_1 = -\left(1 + \frac{\Delta A_c}{A_c}\right)\cos\Delta\theta,\ \beta_2 = 1 + \left(1 + \frac{\Delta A_c}{A_c}\right)\sin\Delta\theta \tag{4.122}$$

$$\beta_1 = -\left(1 + \frac{\Delta A_c}{A_c}\right)\cos\Delta\theta,\ \beta_2 = -1 + \left(1 + \frac{\Delta A_c}{A_c}\right)\sin\Delta\theta. \tag{4.123}$$

The reader is encouraged to compute the error vector magnitude (Chapter 3) of this constellation.

Another approach to quantifying the I/Q mismatch in a transmitter involves applying two tones $V_0\cos\omega_{in}t$ and $V_0\sin\omega_{in}t$ to the I and Q inputs in Fig. 4.88 and examining the output spectrum. In the ideal case, the output is simply given by $V_{out} = V_0\cos\omega_{in}t\cos\omega_c t - V_0\sin\omega_{in}t\sin\omega_c t = V_0\cos(\omega_c + \omega_{in})t$. On the other hand, in the presence of a (relative) gain mismatch of ε and a phase imbalance of $\Delta\theta$, we have

$$V_{out}(t) = V_0(1+\varepsilon)\cos\omega_{in}t\cos(\omega_c t + \Delta\theta) - V_0\sin\omega_{in}t\sin\omega_c t \tag{4.124}$$

$$\begin{aligned} &= \frac{V_0}{2}[(1+\varepsilon)\cos\Delta\theta + 1]\cos(\omega_c t + \omega_{in})t \\ &\quad - \frac{V_0}{2}(1+\varepsilon)\sin\Delta\theta\sin(\omega_c + \omega_{in})t \\ &\quad + \frac{V_0}{2}[(1+\varepsilon)\cos\Delta\theta - 1]\cos(\omega_c - \omega_{in})t \\ &\quad - \frac{V_0}{2}(1+\varepsilon)\sin\Delta\theta\sin(\omega_c - \omega_{in})t. \end{aligned} \tag{4.125}$$

It follows that the power of the unwanted sideband at $\omega_c - \omega_{in}$ divided by that of the wanted sideband at $\omega_c + \omega_{in}$ is given by

$$\frac{P_-}{P_+} = \frac{[(1+\epsilon)\cos\Delta\theta - 1]^2 + (1+\epsilon)^2\sin^2\Delta\theta}{[(1+\epsilon)\cos\Delta\theta + 1]^2 + (1+\epsilon)^2\sin^2\Delta\theta} \tag{4.126}$$

$$= \frac{(1+\epsilon)^2 - 2(1+\epsilon)\cos\Delta\theta + 1}{(1+\epsilon)^2 + 2(1+\epsilon)\cos\Delta\theta + 1}, \tag{4.127}$$

which is similar to the image-rejection ratio expression [Eq. (4.77)]. We may even call the unwanted sideband the "image" of the wanted sideband with respect to the carrier frequency. In practice, a P_-/P_+ of roughly -30 dB is sufficient to ensure negligible distortion of the cancellation, but the exact requirement depends on the type of modulation.

Example 4.38

Compute the average power of $V_{out}(t)$ in Eq. (4.125).

Solution:

We add P_- and P_+ and multiply the result by $V_0^2/4$. If $\epsilon \ll 1$, then

$$\overline{V_{out}^2(t)} = \frac{V_0^2}{2}(1+\epsilon). \tag{4.128}$$

Interestingly, the output power is independent of the phase mismatch.

If the raw I/Q matching of circuits in a transmitter is inadequate, some means of calibration can be employed. To this end, the gain and phase mismatch must first be measured and subsequently corrected. Can we use the power of the unwanted sideband as a symptom of the I/Q mismatch? Yes, but it is difficult to *measure* this power in the presence of the large wanted sideband: the two sidebands are too close to each other to allow suppressing the larger one by tens of decibels by means of a filter.

Let us now apply a single sinusoid to both inputs of the upconverter (Fig. 4.90). The output emerges as

$$V_{out3}(t) = V_0(1+\epsilon)\cos\omega_{in}t\cos(\omega_c t + \Delta\theta) - V_0\cos\omega_{in}t\sin\omega_c t \tag{4.129}$$

$$= V_0\cos\omega_{in}t(1+\epsilon)\cos\Delta\theta\cos\omega_c t$$
$$- V_0\cos\omega_{in}t[(1+\epsilon)\sin\Delta\theta + 1]\sin\omega_c t. \tag{4.130}$$

It can be shown that the output contains two sidebands of equal amplitudes and carries an average power equal to

$$\overline{V_{out3}^2(t)} = V_0^2[1 + (1+\epsilon)\sin\Delta\theta]. \tag{4.131}$$

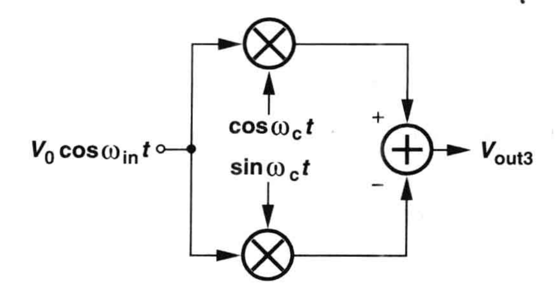


Figure 4.90 *Quadrature upconverter sensing a single sinusoid to reveal phase mismatch.*

We observe that ε is forced to zero as described above, then

$$\overline{V_{out3}^2} - 2\overline{V_{out1}^2} = \sin \Delta\theta. \tag{4.132}$$

Thus, the calibration of phase mismatch proceeds to drive this quantity to zero.

For gain mismatch calibration, we perform two consecutive tests. Depicted in Fig. 4.91, the tests entail applying a sinusoid to one baseband input while the other is set to zero. For the case in Fig. 4.91(a),

$$V_{out1}(t) = V_0(1 + \varepsilon) \cos \omega_{in} t \cos(\omega_c t + \Delta\theta), \tag{4.133}$$

yielding an average power of

$$\overline{V_{out1}^2(t)} = \frac{V_0^2}{2} + V_0^2 \varepsilon. \tag{4.134}$$

In Fig. 4.91(b), on the other hand,

$$V_{out2}(t) = V_0 \cos \omega_{in} t \sin \omega_c t, \tag{4.135}$$

producing

$$\overline{V_{out2}^2(t)} = \frac{V_0^2}{2}. \tag{4.136}$$

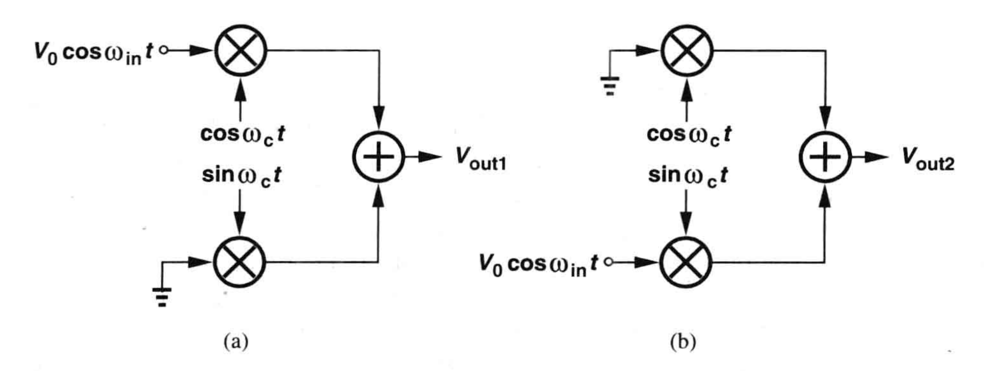


Figure 4.91 *Quadrature upconverter sensing a cosine at (a) the I input, (b) at the Q input.*

That is,

$$\overline{V_{out1}^2(t)} - \overline{V_{out2}^2(t)} = V_0^2\varepsilon, \tag{4.137}$$

suggesting that the gain mismatch can be adjusted so as to drive this difference to zero.

The measurement of $\overline{V_{out3}^2(t)}$ and $\overline{V_{out1}^2(t)} - \overline{V_{out2}^2(t)}$ in the above tests requires a relatively high resolution. For example, a residual phase mismatch of $\Delta\theta = 1^\circ$ translates to $\sin\Delta\theta = 1.75\%$, dictating a resolution of about 7–8 bits in the ADC that digitizes $\overline{V_{out3}^2(t)}$ in Eq. (4.131).

We should remark that dc offsets in the baseband may affect the accuracy of I/Q calibration. As explained below, another effect, "carrier leakage," may also require the removal of dc offsets prior to I/Q calibration.

Carrier Leakage The analog baseband circuitry producing the quadrature signals in the transmitter of Fig. 4.88 exhibits dc offsets, and so does the baseband port of each upconversion mixer. Consequently, the output signal appears as

$$V_{out}(t) = [A(t)\cos\phi + V_{OS1}]\cos\omega_c t - [A(t)\sin\phi + V_{OS2}]\sin\omega_c t, \tag{4.138}$$

where V_{OS1} and V_{OS2} denote the total dc offsets referred to the input port of the mixers. The upconverter output therefore contains a fraction of the *unmodulated* carrier:

$$V_{out}(t) = A(t)\cos(\omega_c t + \phi) + V_{OS1}\cos\omega_c t - V_{OS2}\sin\omega_c t. \tag{4.139}$$

Called "carrier leakage," and quantified as

$$\text{Relative Carrier Leakage} = \frac{\sqrt{V_{OS1}^2 + V_{OS2}^2}}{\sqrt{\overline{A^2(t)}}}, \tag{4.140}$$

this phenomenon leads to two adverse effects. First, it distorts the signal constellation, raising the error vector magnitude at the TX output. For example, if $V_{out}(t)$ represents a QPSK signal,

$$V_{out}(t) = \alpha_1(V_0 + V_{OS1})\cos\omega_c t - \alpha_2(V_0 + V_{OS2})\sin\omega_c t, \tag{4.141}$$

and is applied to an ideal direct-conversion receiver, then the baseband quadrature outputs suffer from dc offsets, i.e., horizontal and vertical shifts in the constellation (Fig. 4.92).

The second effect manifests itself if the output power of the transmitter must be varied across a wide range by varying the amplitude of the baseband signals. For example, as described in Chapter 3, CDMA mobiles must lower their transmitted power as they come closer to the base station so as to avoid the near-far effect. Figure 4.93(a) conceptually depicts the power control loop. The base station measures the power level received from the mobile and, accordingly, requests the mobile to adjust its transmitted power. With a short distance between the two, the mobile output power must be reduced to very low values, yielding the spectrum shown in Fig. 4.93(b) in the presence of carrier leakage. In this case, the carrier power dominates, making it difficult to measure the actual signal power. This

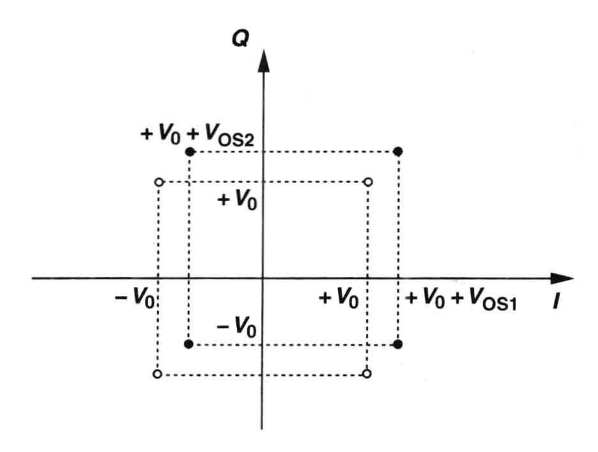


Figure 4.92 *Effect of carrier feedthrough on received signal spectrum.*

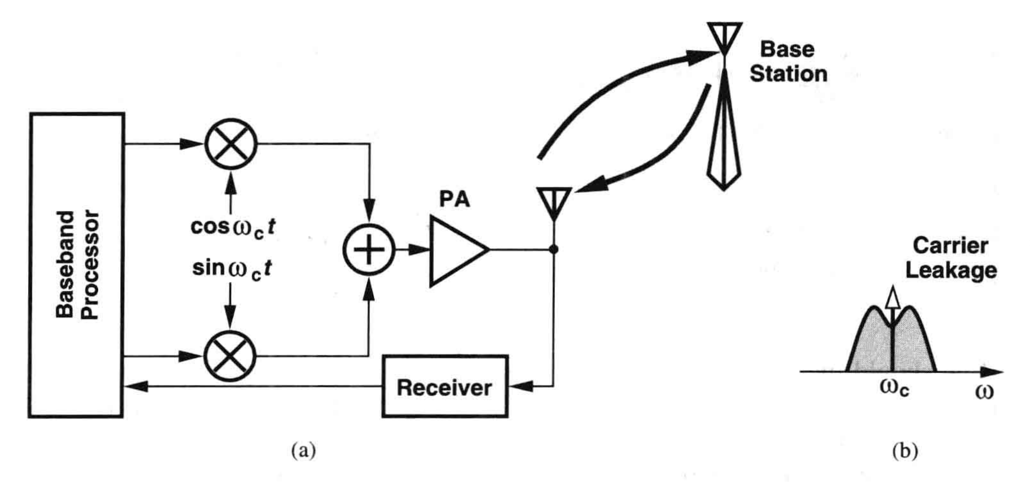


Figure 4.93 *(a) Power control feedback loop in CDMA, (b) effect of carrier leakage.*

issue arises if the mobile output power is adjusted by varying the baseband swings but not if the PA itself is adjusted.

In order to reduce the carrier leakage, Eq. (4.140) suggests that the baseband signal swing, $A(t)$, must be chosen sufficiently large. However, as $A(t)$ increases, the input port of the upconversion mixers becomes more nonlinear. A compromise is therefore necessary. In stringent applications, the offsets must be trimmed to minimize the carrier leakage. As illustrated in Fig. 4.94, two DACs are tied to the baseband ports of the TX[27] and a power detector (e.g., a rectifier or an envelope detector) monitors the output level, and its output is digitized. During carrier leakage cancellation, the baseband processor produces a zero output so that the detector measures only the leakage. Thus, the loop consisting of the TX, the detector, and the DACs drives the leakage toward zero, with the final settings of the DACs stored in the register.

27. The DACs may be embedded within the mixers themselves (Chapter 6).

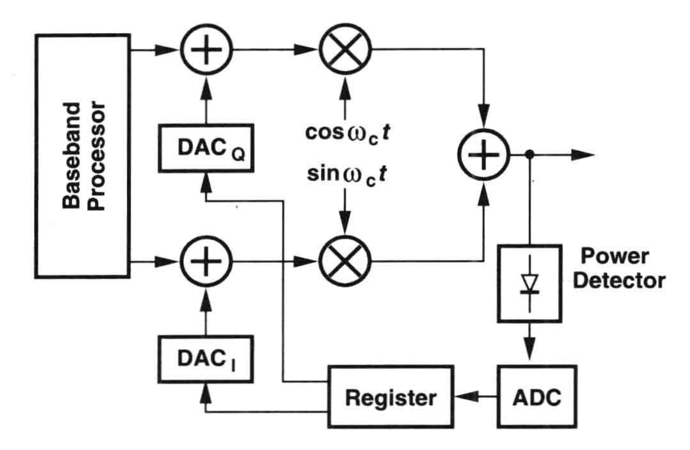


Figure 4.94 *Reduction of carrier leakage by baseband offset control.*

Example 4.39

Is it possible to cancel the carrier leakage by means of a single DAC?

Solution:

No, it is not. Eq. (4.139) implies that no choice of V_{OS1} or V_{OS2} can force $V_{OS1}\cos\omega_c t - V_{OS2}\sin\omega_c t$ to zero if the other remains finite.

How should the two DACs be adjusted so that the loop in Fig. 4.94 converges? An adaptive loop such as the least mean square (LMS) algorithm can perform this task. Alternatively, an "exhaustive" search can arrive at the optimum settings; e.g., for 8-bit DACs, only 256×256 possible combinations exist, and the system can try all to determine which combination yields the lowest leakage at the output. In this procedure, the system begins with, say, zero settings for both DACs, measures the carrier leakage, *memorizes* this value, increments one DAC by 1 LSB, measures the leakage again, and compares the result with the previous value. The new settings replace the previous ones if the new leakage is lower.

Mixer Linearity Unlike downconversion mixers in a receiver, upconversion mixers in a transmitter sense no interferers. However, excessive nonlinearity in the baseband port of upconversion mixers can corrupt the signal or raise the adjacent channel power (Chapter 3). As an example, consider the GMSK signal expressed by Eq. (4.112) and suppose the baseband I/Q inputs experience a nonlinearity given by $\alpha_1 x + \alpha_3 x^3$. The upconverted signal assumes the form [15]

$$V_{out}(t) = (\alpha_1 A\cos\phi + \alpha_3 A^3\cos^3\phi)\cos\omega_c t - (\alpha_1 A\sin\phi + \alpha_3 A^3\sin^3\phi)\sin\omega_c t \quad (4.142)$$

$$= \left(\alpha_1 A + \frac{3}{4}\alpha_3 A^3\right)\cos(\omega_c t + \phi) + \frac{\alpha_3 A^3}{4}\cos(\omega_c t - 3\phi). \quad (4.143)$$

The second term also represents a GMSK signal but with a threefold modulation index, thereby occupying a larger bandwidth. Simulations indicate that this effect becomes negligible if the baseband port of the mixers experiences a nonlinearity less than 1% for the specified baseband swings [15].

For variable-envelope signals, $A^3(t)$ appears in both terms of Eq. (4.143), exacerbating the effect. The required mixer linearity is typically determined by simulations. However, in most cases (i.e., in a good design), as the baseband signal swings increase, the PA output begins to compress *before* the mixer nonlinearity manifests itself. This is explained below.

TX Linearity The linearity of transmitters must be chosen according to the spectral regrowth (adjacent channel power) requirements and/or the tolerable distortion of the signal to be transmitted. As mentioned in Chapter 3, both of these effects become critical for variable-envelope modulation schemes. We defer the former to Chapter 12 and deal with the latter here.

The distortion of a variable-envelope signal is typically characterized by the compression that it experiences. As shown in Fig. 4.95, the signal is subjected to a nonlinear characteristic in simulations and it is determined how close its level can come to the 1-dB compression point while the degradation in the constellation or the bit error rate is negligible. For example, the average power of the 64-QAM OFDM signal in 802.11a must remain about 8 dB below P_{1dB} of a given circuit. We say the circuit must operate at "8-dB back-off." In other words, if a peak swing of V_0 places the circuit at the 1-dB compression point, then the average signal swing must not exceed $V_0/2.51$.

In a TX chain, the signal may experience compression in any of the stages. Consider the example depicted in Fig. 4.96, where the signal levels (in dB) along the chain are also shown. Since the largest voltage swing occurs at the output of the PA, this stage dominates the compression of the TX; i.e., in a good design, the preceding stages must remain well below compression as the PA output approaches P_{1dB}. To so ensure, we must *maximize* the gain of the PA and minimize the output swing of the predriver and the stages preceding it. This requirement places additional burden on the PA design (Chapters 12 and 13).

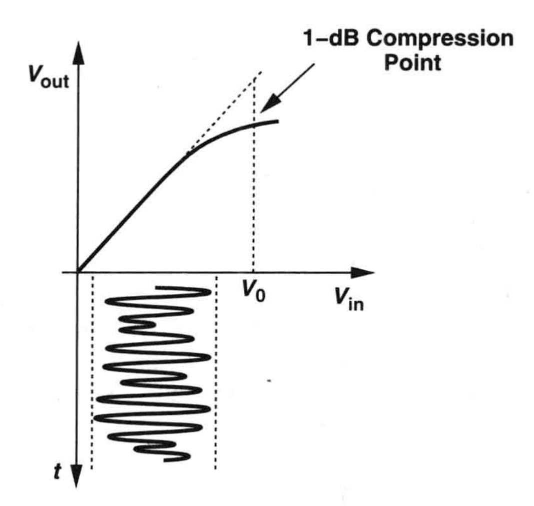


Figure 4.95 *Variable-envelope signal applied to a compressive system.*

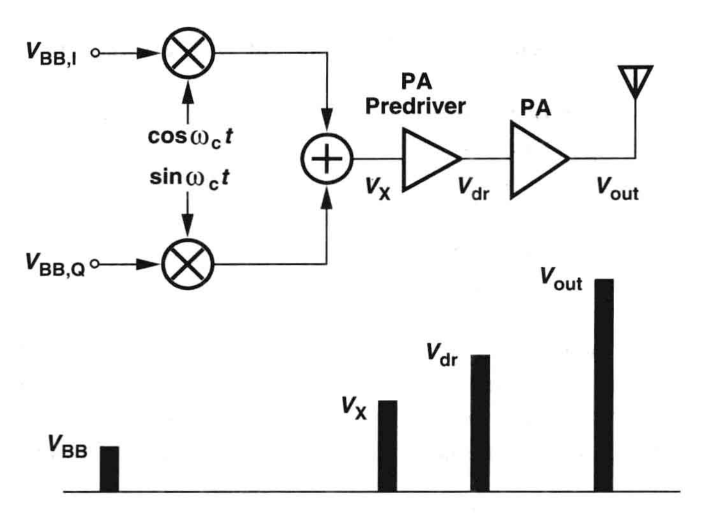


Figure 4.96 *Level chart for the signals along a TX chain.*

Example 4.40

If the predriver and the PA exhibit third-order characteristics, compute the 1-dB compression point of the cascade of the two.

Solution:

Assuming a nonlinearity of $\alpha_1 x + \alpha_3 x^3$ for the predriver and $\beta_1 x + \beta_3 x^3$ for the PA, we write the PA output as

$$y(t) = \beta_1(\alpha_1 x + \alpha_3 x^3) + \beta_3(\alpha_1 x + \alpha_3 x^3)^3 \tag{4.144}$$

$$= \beta_1\alpha_1 x + (\beta_1\alpha_3 + \beta_3\alpha_1^3)x^3 + \cdots \tag{4.145}$$

If the first two terms are dominant, then the input 1-dB compression point is given by the coefficients of x and x^3 as follows:

$$A_{1dB,in} = \sqrt{0.145\left|\frac{\beta_1\alpha_1}{\beta_1\alpha_3 + \beta_3\alpha_1^3}\right|}. \tag{4.146}$$

The reader is encouraged to consider the special cases $\beta_3 = 0$ or $\alpha_3 = 0$ and justify the results intuitively. It is interesting to note that if $\beta_1\alpha_3 = -\beta_3\alpha_1^3$, the coefficient of x^3 falls to zero and $A_{1dB,in} \to \infty$. This is because the compression in one stage is cancelled by the expansion in the other.

In transmitters, the *output* power is of interest, suggesting that the compression behavior must also be quantified at the output. Since at $A_{1dB,in}$, the output level is 1 dB below its ideal value, we simply multiply $A_{1dB,in}$ by the total gain and reduce the result by 1 dB so as to determine the output compression point:

$$A_{1dB,out} = A_{1dB,in} \times |\alpha_1\beta_1| \times \frac{1}{1.12}, \tag{4.147}$$

Example 4.40 *(Continued)*

where the factor 1/1.12 accounts for the 1-dB gain reduction. It follows that

$$A_{1dB,out} = \frac{0.34|\alpha_1\beta_1|\sqrt{|\beta_1\alpha_1|}}{\sqrt{\beta_1\alpha_3 + \beta_3\alpha_1^3}}. \tag{4.148}$$

Oscillator Pulling While the issues described above apply to most transmitter architectures, another one becomes particularly critical in direct-conversion topologies. As illustrated in Fig. 4.97(a), the PA output exhibits very large swings (20 V_{pp} for 1 W delivered to a 50-Ω load), which couple to various parts of the system through the silicon substrate, package parasitics, and traces on the printed-circuit board. Thus, it is likely that an appreciable fraction of the PA output couples to the local oscillator. Even if the PA is off-chip, the PA driver may still pull the LO. Note that the center frequency of the PA output spectrum is equal to ω_{LO} in direct-conversion transmitters.

Let us consider a "sliver" of the output spectrum centered at $\omega_{LO} + \Delta\omega$ and model it by an impulse of equal energy [Fig. 4.97(b)]. We therefore inquire what happens if a sinusoid at a frequency of $\omega_1 = \omega_{LO} + \Delta\omega$ is "injected" into an oscillator operating at a frequency of ω_{LO}, where $\Delta\omega \ll \omega_{LO}$. Called "injection pulling," this phenomenon has been studied extensively [16, 17] and is analyzed in Chapter 8. In such a condition, the output phase of the oscillator, ϕ_{out}, is modulated *periodically*. In fact, as depicted in Fig. 4.98(a), ϕ_{out} remains around 90° (with respect to the input phase) for part of the period, subsequently experiencing a rapid 360° rotation. The input and output waveforms therefore appear as in Fig. 4.98(b). It can be proved that the output spectrum is heavily asymmetric [Fig. 4.98(c)], with most of the impulses located *away* from the input frequency, ω_{inj} $(=\omega_{LO} + \Delta\omega)$. Note that the spacing between the impulses is equal to the frequency of the phase variation in Fig. 4.98(a) and *not* equal to $\Delta\omega$.[28]

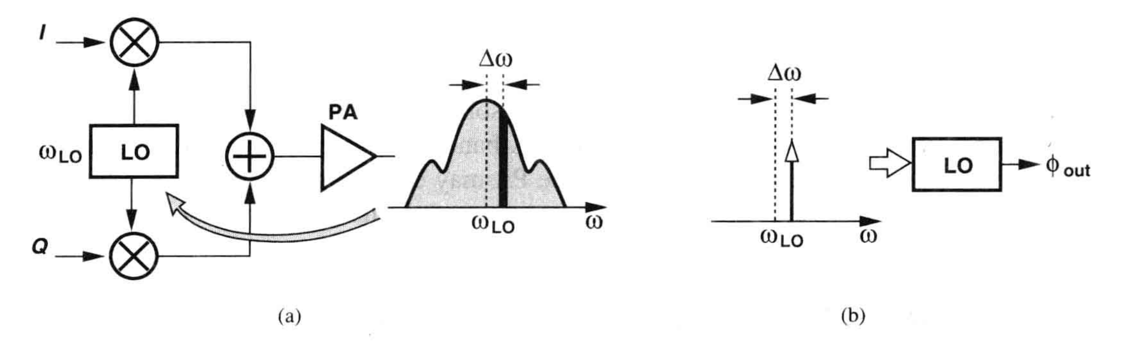


Figure 4.97 *(a) Injection pulling of LO by PA in a direct-conversion TX, (b) conceptual illustration of injection into an oscillator.*

28. Pulling can also occur if the injected signal frequency is close to a *harmonic* of the oscillator frequency, e.g., in the vicinity of $2\omega_{LO}$. We call this effect "superharmonic pulling."

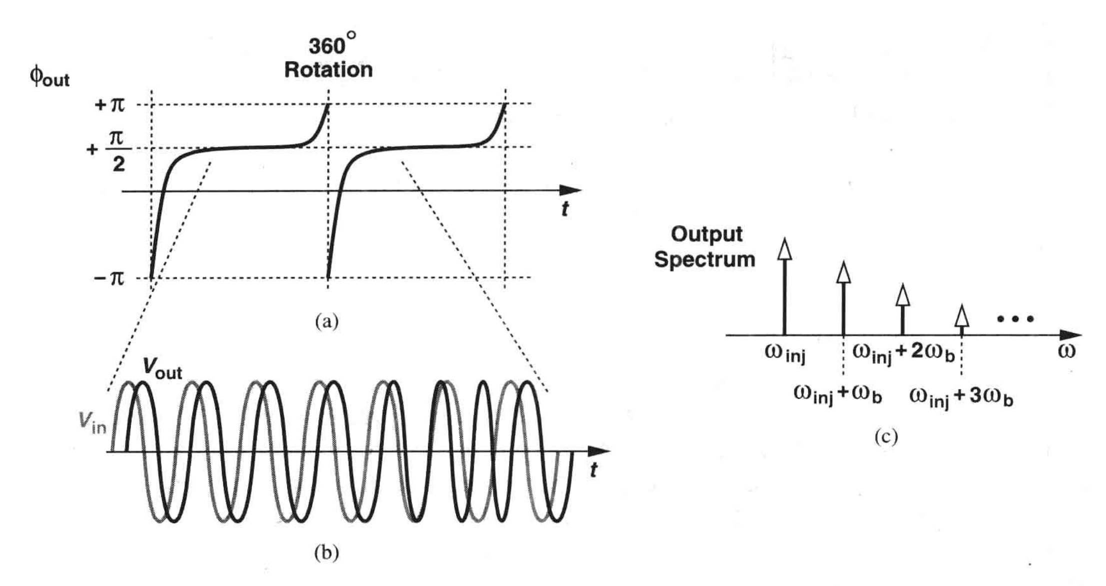


Figure 4.98 *(a) Behavior of LO output phase in the presence of injection pulling, (b) cycle slips in time domain, (c) resulting spectrum.*

At what output power level does injection pulling become significant? The answer depends on several factors: (1) the internal voltage and current swings of the LO (the larger they are, the less is the effect of pulling); (2) the Q of the tank used in the oscillator; (3) whether the PA output is differential, in which case its coupling to the LO is 30–40 dB lower than if it is single-ended;[29] (4) how much the feedback loop controlling the LO (the synthesizer) counteracts the pulling [17]; (5) the symmetry of the layout and the type of packaging. Nonetheless, for typical designs, as the PA output exceeds 0 dBm, injection pulling may prove serious.

In order to avoid injection pulling, the PA output frequency and the oscillator frequency must be made sufficiently different (e.g., by more than 20%), an impossible task in the architecture of Fig. 4.97. This principle has led to a number of transmitter architectures and frequency plans that are described below.

Noise in RX Band As explained in Chapter 3, some standards (e.g., GSM) specify the maximum noise that a TX can transmit in the RX band. In a direct-conversion transmitter, the baseband circuits, the upconverter, and the PA may create significant noise in the RX band. To resolve this issue, "offset-PLL" transmitters can be used (Chapter 9).

4.3.3 Modern Direct-Conversion Transmitters

Most of today's direct-conversion transmitters avoid an oscillator frequency equal to the PA output frequency. To avoid confusion, we call the former the LO frequency, ω_{LO}, and the

29. This is true only if the differential PA incorporates "single-ended" inductors rather than one symmetric inductor (Chapter 7).

latter, the carrier frequency, ω_c. The task is accomplished by choosing ω_{LO} sufficiently far from ω_c and *deriving* ω_c from ω_{LO} by operations such as frequency division and mixing.

Figure 4.99 depicts a common example where $\omega_{LO} = 2\omega_c$. A divide-by-2 circuit follows the LO, thereby generating $\omega_{LO}/2$ with quadrature phases. This architecture is popular for two reasons: (1) injection pulling is greatly reduced, and (2) the divider readily provides quadrature phases of the carrier, an otherwise difficult task (Chapter 8).

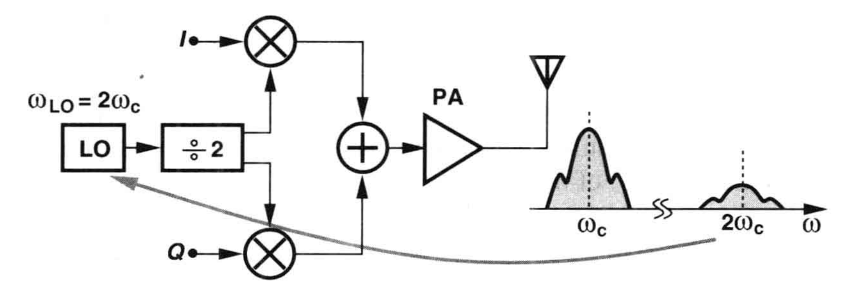


Figure 4.99 *Use of an LO running at twice the carrier frequency to minimize LO pulling.*

The architecture of Fig. 4.99 does not entirely eliminate injection pulling. Since the PA nonlinearity produces a finite amount of power at the *second* harmonic of the carrier, the LO may still be pulled. Nonetheless, proper layout and isolation techniques can suppress this effect.

Example 4.41

Is it possible to choose $\omega_{LO} = \omega_c/2$ and use a frequency *doubler* to generate ω_c?

Solution:

It is possible, but the doubler typically does not provide quadrature phases, necessitating additional quadrature generation stages. Figure 4.100 shows an example where the doubler output is applied to a polyphase filter (Section 4.2.5). The advantage of this architecture is that no harmonic of the PA output can pull the LO. The serious disadvantage is that the doubler and the polyphase filter suffer from a high loss, requiring the use of power-hungry buffers.

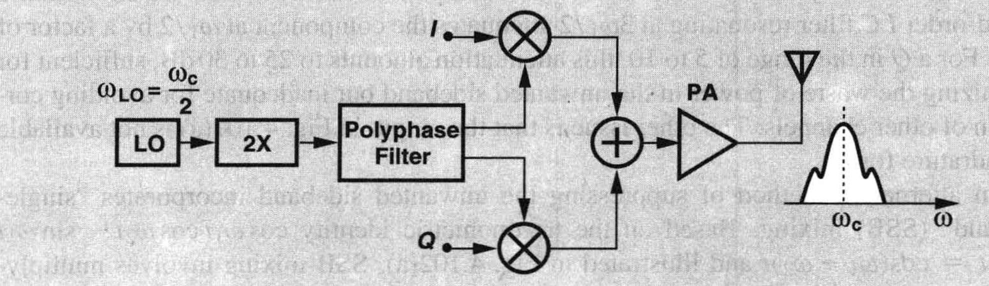


Figure 4.100 *Use of an LO running at half the carrier frequency to minimize LO pulling.*

The principal drawback of the architecture of Fig. 4.99 stems from the speed required of the divider. Operating at twice the carrier frequency, the divider may become the speed bottleneck of the entire transceiver. (We deal with divider design in Chapter 10.) Nevertheless, as seen in the following discussion, other transmitter architectures suffer from more serious drawbacks. Thus, even a substantial effort on divider design to enable this architecture is well justified.

Another approach to deriving frequencies is through the use of mixing. For example, mixing the outputs of two oscillators operating at ω_1 and ω_2 can yield $\omega_1 + \omega_2$ or $\omega_1 - \omega_2$. Nonetheless, as with the receivers studied in Section 4.2, it is desirable to employ a *single* oscillator and utilize division to obtain subharmonics. To this end, let us consider the arrangement shown in Fig. 4.101(a), where the oscillator frequency is divided by 2 and the two outputs are mixed. The result contains components at $\omega_1 \pm \omega_1/2$ with equal magnitudes. We may call one the "image" of the other with respect to ω_1.

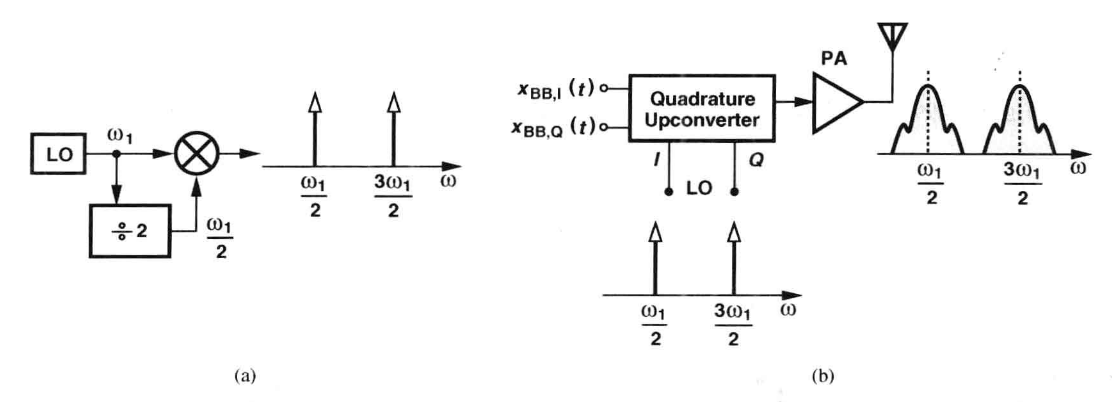


Figure 4.101 *(a) Mixing of an LO output with half of its frequency, (b) effect of two sidebands on transmitter output.*

Can both components be retained? In a transmitter using such an LO waveform, the upconverter output would contain, with equal power, the signal spectrum at both carrier frequencies [Fig. 4.101(b)]. Thus, half of the power delivered to the antenna is wasted. Furthermore, the power transmitted at the unwanted carrier frequency corrupts communication in other channels or bands. One component (e.g., that at $\omega_1/2$) must therefore be suppressed.

It is difficult to remove the unwanted carrier by means of filtering because the two frequencies differ by only a factor of 3. For example, in Problem 4.23 we show that a second-order *LC* filter resonating at $3\omega_1/2$ attenuates the component at $\omega_1/2$ by a factor of $8Q/3$. For a Q in the range of 5 to 10, this attenuation amounts to 25 to 30 dB, sufficient for minimizing the waste of power in the unwanted sideband but inadequate for avoiding corruption of other channels. The other issue is that the output in Fig. 4.101(a) is not available in quadrature form.

An alternative method of suppressing the unwanted sideband incorporates "single-sideband" (SSB) mixing. Based on the trigonometric identity $\cos\omega_1 t \cos\omega_2 t - \sin\omega_1 t \sin\omega_2 t = \cos(\omega_1 + \omega_2)t$ and illustrated in Fig. 4.102(a), SSB mixing involves multiplying the quadrature phases of ω_1 and ω_2 and subtracting the results—just as realized by the quadrature upconverter of Fig. 4.88. We denote an SSB mixer by the abbreviated symbol

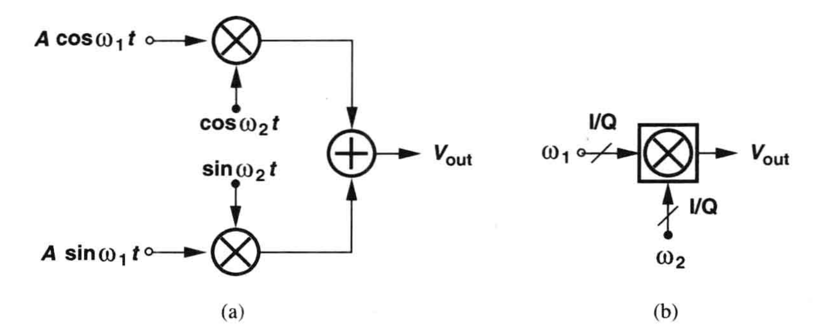


Figure 4.102 *Single-sideband mixer (a) implementation, (b) simplified symbol.*

shown in Fig. 4.102(b). Of course, gain and phase mismatches lead to an unwanted sideband, as expressed by Eq. (4.127). With typical mismatches, P_-/P_+ falls in the vicinity of 30 to 40 dB, and filtering by a second-order *LC* stage attenuates the sideband by another 25 to 30 dB.

In addition to the image sideband, the *harmonics* of the input frequencies also corrupt the output of an SSB mixer. For example, suppose each mixer in Fig. 4.102(a) exhibits third-order nonlinearity in the port sensing $A \sin \omega_1 t$ or $A \cos \omega_1 t$. If the nonlinearity is of the form $\alpha_1 x + \alpha_3 x^3$, the output can be expressed as

$$V_{out}(t) = (\alpha_1 A \cos \omega_1 t + \alpha_3 A^3 \cos^3 \omega_1 t) \cos \omega_2 t - (\alpha_1 A \sin \omega_1 t + \alpha_3 A^3 \sin^3 \omega_1 t) \sin \omega_2 t \tag{4.149}$$

$$= \left(\alpha_1 A + \frac{3\alpha_3 A^3}{4}\right) \cos \omega_1 t \cos \omega_2 t - \left(\alpha_1 A + \frac{3\alpha_3 A^3}{4}\right) \sin \omega_1 t \sin \omega_2 t + \frac{\alpha_3 A^3}{4} \cos 3\omega_1 t \cos \omega_2 t + \frac{\alpha_3 A^3}{4} \sin 3\omega_1 t \sin \omega_2 t \tag{4.150}$$

$$= \left(\alpha_1 A + \frac{3\alpha_3 A^3}{4}\right) \cos(\omega_1 + \omega_2)t + \frac{\alpha_3 A^3}{4} \cos(3\omega_1 - \omega_2)t. \tag{4.151}$$

The output spectrum contains a spur at $3\omega_1 - \omega_2$. Similarly, with third-order nonlinearity in the mixer ports sensing $\sin \omega_2 t$ and $\cos \omega_2 t$, a component at $3\omega_2 - \omega_1$ arises at the output. The overall output spectrum (in the presence of mismatches) is depicted in Fig. 4.103.

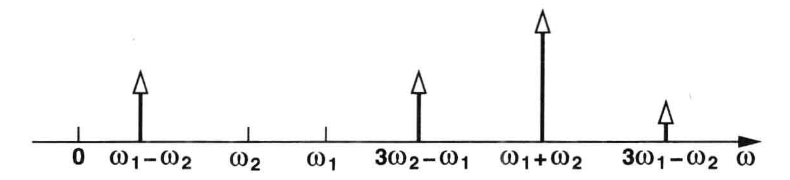


Figure 4.103 *Output spectrum of SSB mixer in the presence of nonlinearity and mismatches.*

Figure 4.104 shows a mixer example where the port sensing V_{in1} is linear while that driven by V_{in2} is nonlinear. As explained in Chapter 2, the circuit multiplies V_{in1} by a *square wave* toggling between 0 and 1. That is, the third harmonic of V_{in2} is only one-third of its fundamental, thus producing the strong spurs in Fig. 4.103.

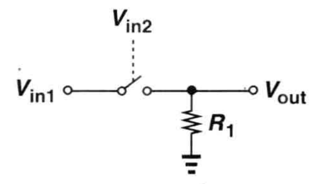


Figure 4.104 *Simple mixer.*

How serious are the above spurs? In a typical mixer design (Chapter 6), it is possible to linearize only one port, thus maintaining a small third harmonic in that port. The other is highly nonlinear so as to retain a reasonable gain (or loss). We therefore conclude that, between the two spurs at $3\omega_1 - \omega_2$ and $3\omega_2 - \omega_1$, only one can be reduced to acceptably low levels while the other remains only 10 dB (a factor of one-third) below the desired component. As an example, if $\omega_2 = \omega_1/2$, then $3\omega_1 - \omega_2 = 5\omega_1/2$ and $3\omega_2 - \omega_1 = \omega_1/2$; we can linearize the port sensing ω_2 to suppress the latter, but the former still requires substantial filtering.

For use in a direct-conversion TX, the SSB mixer must provide the quadrature phases of the carrier. This is accomplished by noting that $\sin\omega_1 t \cos\omega_2 t + \cos\omega_1 t \sin\omega_2 t = \sin(\omega_1 + \omega_2)t$ and duplicating the SSB mixer as shown in Fig. 4.105.

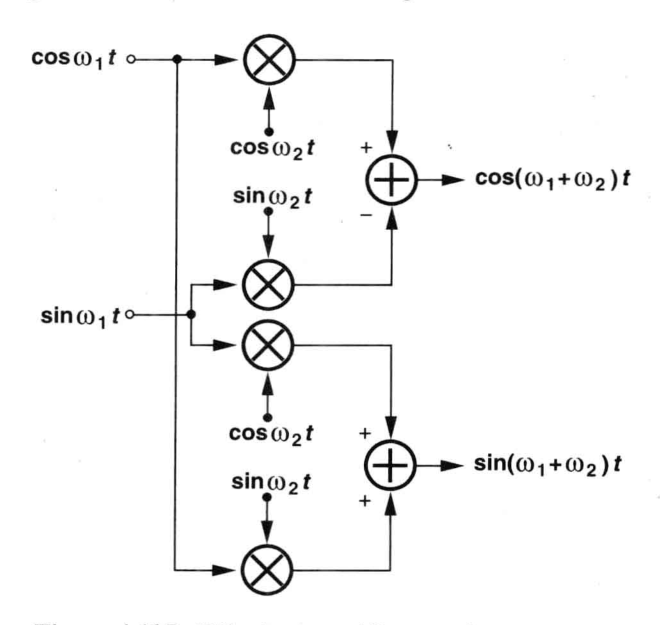


Figure 4.105 *SSB mixer providing quadrature outputs.*

Figure 4.106 shows a direct-conversion TX with SSB mixing for carrier generation. Since the carrier and LO frequencies are sufficiently different, this architecture remains free from injection pulling.[30] While suppressing the carrier sideband at $\omega_1/2$, this architecture

30. This is not strictly correct because the second harmonic of the PA output is also the third harmonic of the LO, potentially causing "superharmonic" pulling.

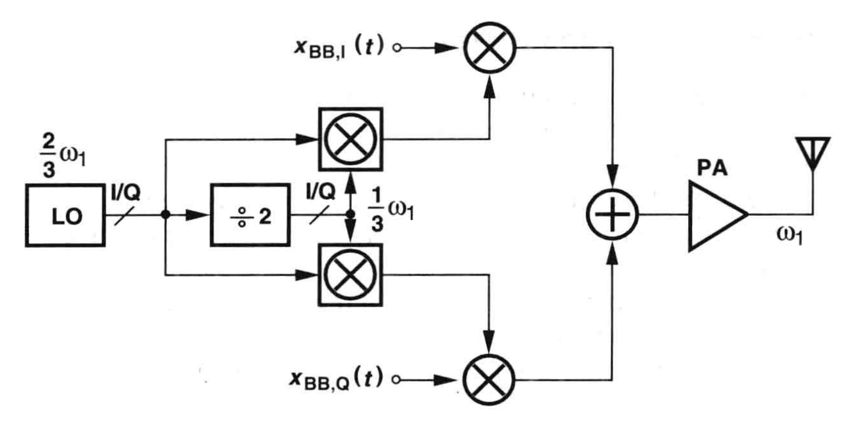


Figure 4.106 *Direct-conversion TX using SSB mixing in LO path.*

presents two drawbacks: (1) the spurs at $5\omega_1/2$ and other harmonic-related frequencies prove troublesome, and (2) the LO must provide quadrature phases, a difficult issue (Chapter 8).

Example 4.42

A student replaces the ÷2 circuit in Fig. 4.106 with a ÷4 topology. Analyze the unwanted components in the carrier.

Solution:

Upon mixing ω_1 and $\omega_1/4$, the SSB mixer generates $5\omega_1/4$ and, due to mismatches, $3\omega_1/4$. In the previous case, these values were given by $3\omega_1/2$ and $\omega_1/2$, respectively. Thus, filtering the unwanted sideband is more difficult in this case because it is closer to the wanted sideband.

As for the effect of harmonics, the output contains spurs at $3\omega_1 - \omega_2$ and $3\omega_2 - \omega_1$, which are respectively equal to $11\omega_1/4$ and $\omega_1/4$ if $\omega_2 = \omega_1/4$. The spur at $11\omega_1/4$ remains slightly higher than its counterpart in the previous case ($5\omega_1/2$), while that at $\omega_1/4$ is substantially lower and can be filtered more easily. Figure 4.107 summarizes the output components.

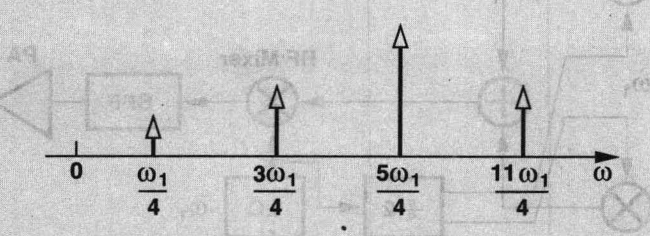


Figure 4.107 *Output spurs with a divide-by-4 circuit used in LO mixing.*

4.3.4 Heterodyne Transmitters

Another approach to avoiding injection pulling involves performing the signal upconversion in *two* steps so that the LO frequency remains far from the PA output spectrum. Shown in Fig. 4.108, such a TX topology is the "dual" of the heterodyne receiver studied in Section 4.2.1. Here, the baseband I and Q signals are upconverted to an IF of ω_1, and the result is mixed with ω_2 and hence translated to a carrier frequency of $\omega_1 + \omega_2$. Since the second mixer also produces an output at $\omega_1 - \omega_2$, a band-pass filter follows this stage. As with the receiver counterpart, one advantage of this architecture is that the I/Q upconversion occurs at a significantly lower frequency than the carrier, exhibiting smaller gain and phase mismatches. The equations quantifying the effect of mismatches are the same as those derived in Section 4.3.2 for the direct-conversion TX.

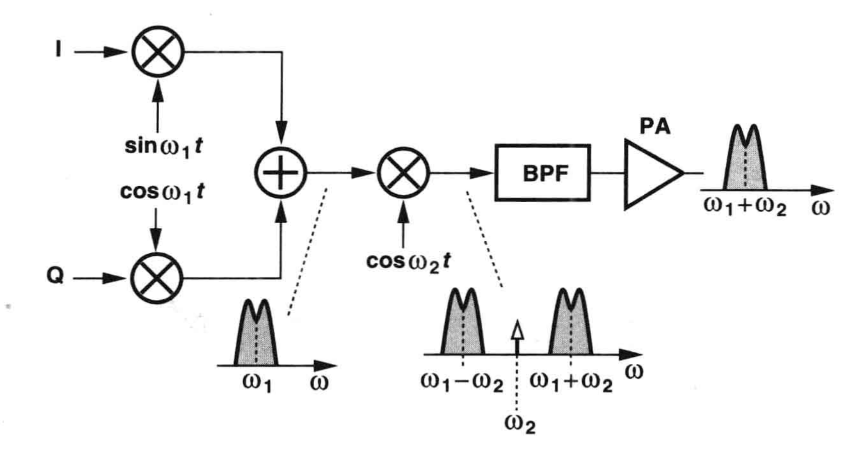


Figure 4.108 *Two-step TX.*

In analogy with the sliding-IF receiver architecture of Fig. 4.26(a), we eliminate the first oscillator in the above TX and derive the required phases from the second oscillator (Fig. 4.109). The carrier frequency is thus equal to $3\omega_1/2$. Let us study the effect of nonidealities in this architecture. We call the LO waveforms at $\omega_1/2$ and ω_1 the first and second LOs, respectively.

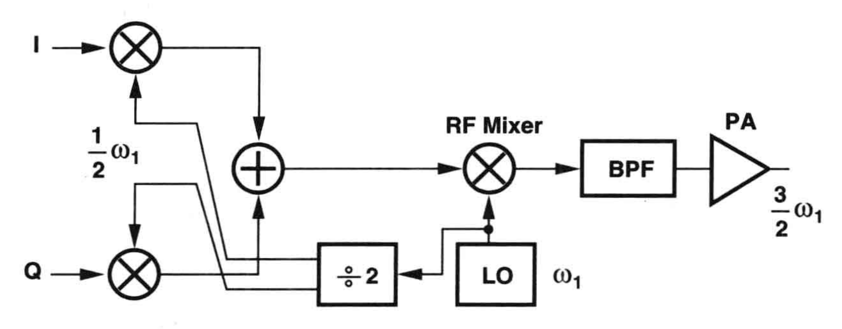


Figure 4.109 *Sliding-IF TX.*

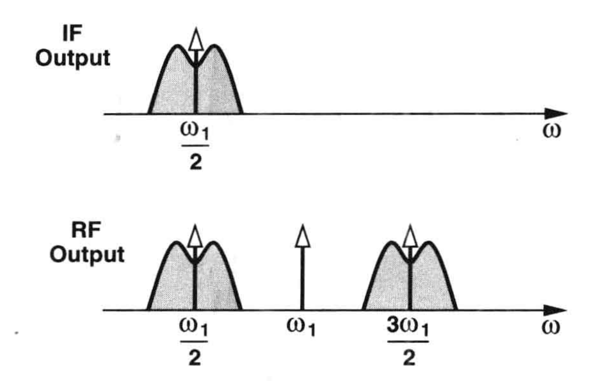


Figure 4.110 *Carrier leakage in heterodyne TX.*

Carrier Leakage The dc offsets in the baseband yield a component at $\omega_1/2$ at the output of the quadrature upconverter, and the dc offset at the input of the RF mixer produces another component at ω_1 (Fig. 4.110). The former can be minimized as described in Section 4.3.2. The latter, and the lower sideband at $\omega_1/2$, must be removed by filtering. The leakage at ω_1 is closer to the upper sideband than the lower sideband is, but it is also much smaller than the lower sideband. Thus, the filter following the RF mixer must be designed to attenuate both to acceptably low levels.

Mixing Spurs The heterodyne TX of Fig. 4.109 displays various mixing spurs that must be managed properly. The spurs arise from two mechanisms: the harmonics of the first LO and the harmonics of the second LO.

The quadrature upconverter mixes the baseband signals with the third and fifth harmonics of the first LO,[31] thus creating replicas of the signal spectrum at $\omega_1/2$, $3\omega_1/2$, and $5\omega_1/2$. The result is shown in Fig. 4.111(a) for an asymmetrically-modulated signal. Note that the harmonic magnitudes follow a sinc envelope if the mixers operate as the switching network depicted in Fig. 4.104. In other words, the magnitudes of the replicas at $3\omega_1/2$ and $5\omega_1/2$ are one-third and one-fifth of the desired signal magnitude, respectively. Upon mixing with the second LO (ω_1), the components in Fig. 4.111(a) are translated up and down by an amount equal to ω_1, yielding the spectrum illustrated in Fig. 4.111(b). Interestingly, the desired sideband at $+3\omega_1/2$ is *enhanced* by a smaller replica that results from the mixing of $5\omega_1/2$ and ω_1. The unwanted sidebands at $\omega_1/2$, $5\omega_1/2$, and $7\omega_1/2$ must be suppressed by an RF band-pass filter.

The second mechanism relates to the harmonics of the second LO. That is, the spectrum shown in Fig. 4.111(a) is mixed with not only ω_1 but $3\omega_1$, $5\omega_1$, etc. Illustrated in Fig. 4.112 is the resulting output, revealing that, upon mixing with $+3\omega_1$, the IF sideband at $-3\omega_1/2$ is translated to $+3\omega_1/2$, thereby *corrupting* the wanted sideband (if the modulation is asymmetric). Similarly, the IF sideband at $-5\omega_1/2$ is mixed with $+5\omega_1$ and falls atop the desired signal.

31. The higher harmonics are neglected here.

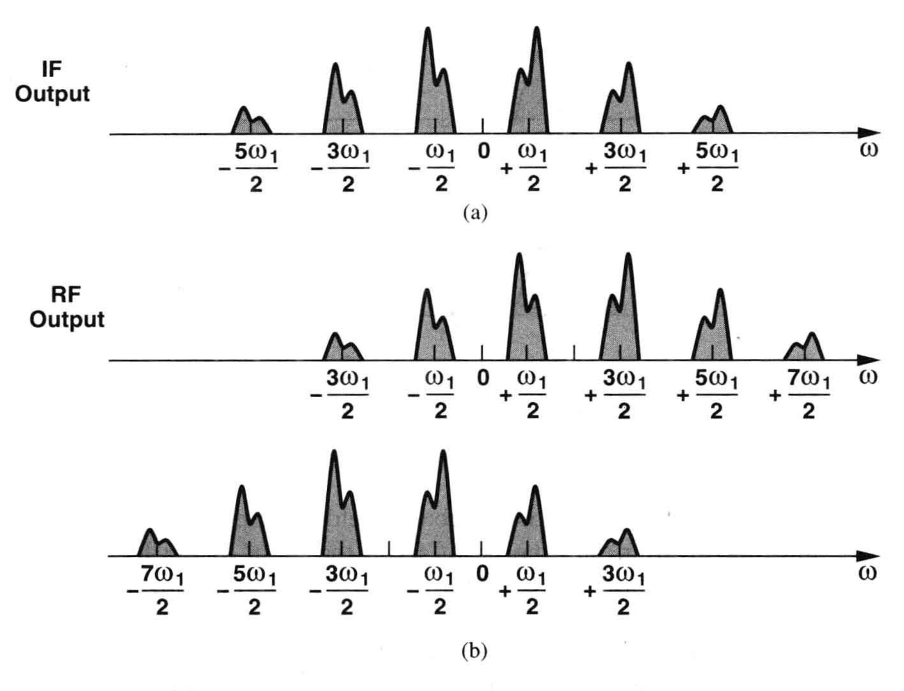


Figure 4.111 *Spurs at (a) IF and (b) RF outputs of a heterodyne TX.*

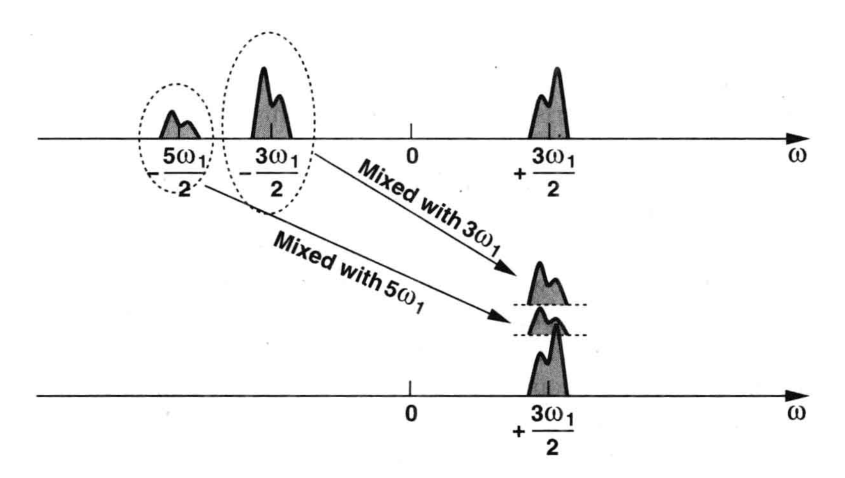


Figure 4.112 *Effect of harmonics of second LO on TX output.*

How serious is this corruption? Since the IF sideband at $-3\omega_1/2$ is 10 dB below the desired signal, and since mixing with $3\omega_1$ entails another 10 dB attenuation (why?), the level of corruption is at -20 dB. This value is acceptable only for modulation schemes that require a moderate SNR (10–12 dB) (e.g., QPSK) or systems with a moderate bit error rate (e.g., 10^{-2}). Even in these cases, some IF filtering is necessary to suppress the unwanted sidebands before they are upconverted to RF and fall into other users' channels.

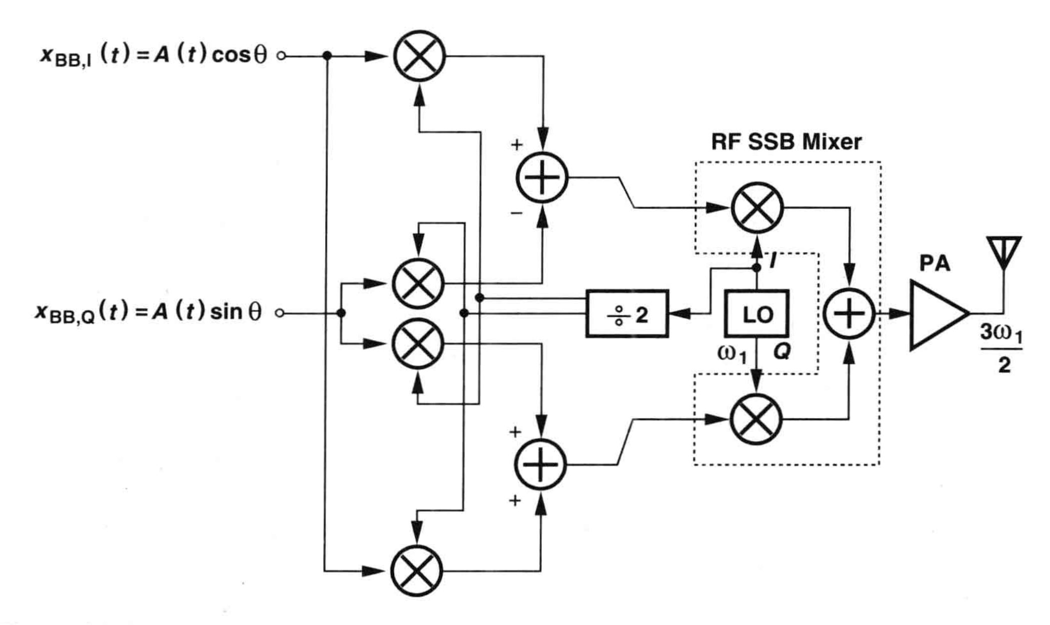


Figure 4.113 *Use of baseband quadrature SSB mixing and IF SSB mixing to reduce the unwanted component.*

Example 4.43

Compare the spurious behavior of the TX architectures shown in Figs. 4.106 and 4.109.

Solution:

In the direct-conversion TX of Fig. 4.106, the primary spur appears at $5\omega_1/2$, and no self-corruption similar to that illustrated in Fig. 4.112 exists. The heterodyne topology, on the other hand, suffers from more spurs.

The unwanted sideband at $\omega_1 - \omega_1/2$ produced by the RF mixer in Fig. 4.109 can be greatly suppressed through the use of SSB mixing. To this end, the IF signal must be generated in *quadrature* form. Figure 4.113 shows such a topology [15, 18], where two quadrature upconverters provide the quadrature components of the IF signal:

$$x_{IF,I}(t) = A(t)\cos\theta\cos\frac{\omega_1 t}{2} - A(t)\sin\theta\sin\frac{\omega_1 t}{2} \tag{4.152}$$

$$= A(t)\cos\left(\frac{\omega_1 t}{2} + \theta\right) \tag{4.153}$$

$$x_{IF,Q}(t) = A(t)\cos\theta\sin\frac{\omega_1 t}{2} + A(t)\sin\theta\cos\frac{\omega_1 t}{2} \tag{4.154}$$

$$= A(t)\sin\left(\frac{\omega_1 t}{2} + \theta\right). \tag{4.155}$$

The RF SSB mixer then translates the result to $\omega_1 + \omega_1/2$. The reader is encouraged to study the mixing spurs in this architecture.

While attenuating the sideband at $\omega_1 - \omega_1/2$, the architecture of Fig. 4.113 suffers from three drawbacks: (1) the oscillator must provide quadrature outputs, a difficult issue (Chapter 8), (2) the circuit employs twice as many mixers as those in the original architecture (Fig. 4.109), and (3) the loading seen by the $\div 2$ circuit is doubled. The first issue can be alleviated by operating the oscillator at $2\omega_1$ and following it with a $\div 2$ stage, but such a design is only slightly simpler than the direct-conversion architecture of Fig. 4.106.

Our study of the heterodyne sliding-IF TX has thus far assumed that the first LO frequency is half of the second LO frequency. It is possible to replace the $\div 2$ circuit with a $\div 4$ stage so as to produce the IF signal at $\omega_1/4$ and the RF output at $\omega_1 + \omega_1/4 = 5\omega_1/4$ [19]. We study the spurious effects in such an architecture in Problem 4.24.

4.3.5 Other TX Architectures

In addition to the TX architectures described above, several others find usage in some applications. These include "offset PLL" topologies, "in-loop modulation" systems, and "polar modulation" transmitters. We study the first two in Chapter 10 and the last in Chapter 12.

4.4 OOK TRANSCEIVERS

"On-off keying" (OOK) modulation is a special case of ASK where the carrier amplitude is switched between zero and maximum. Transceivers employing OOK lend themselves to a compact, low-power implementation and merit some study here. Figure 4.114 illustrates two TX topologies. In Fig. 4.114(a), the LO is directly turned on and off by the binary baseband data. If the LO swings are large enough, the PA also experiences relatively complete switching and delivers an OOK waveform to the antenna. In contrast to the transmitter architectures studied in the previous sections, OOK does not require quadrature baseband or LO waveforms or a quadrature upconverter. Of course, it is also less bandwidth-efficient as unshaped binary pulses modulated on one phase of the carrier occupy a wide spectrum.

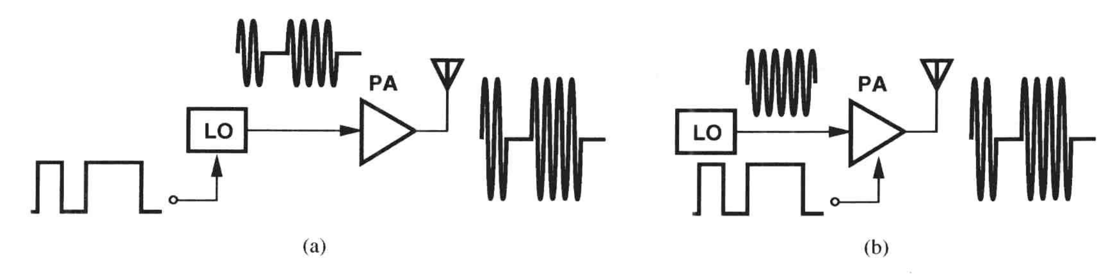


Figure 4.114 *OOK TX with (a) direct LO switching (b) PA switching.*

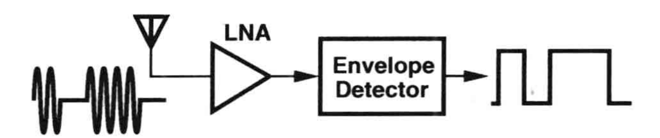


Figure 4.115 *OOK receiver.*

Nonetheless, the simplicity of the architecture makes it attractive for low-cost, low-power applications.

The principal issue in the TX of Fig. 4.114(a) is that the LO cannot be easily controlled by a phase-locked loop (Chapter 9). The TX of Fig. 4.114(b), on the other hand, keeps the LO on and directly switches the PA. We study the injection-pulling properties of the two architectures in Problem 4.28.

OOK receivers are also simple and compact. As shown in Fig. 4.115, an LNA followed by an envelope detector can recover the binary data, with no need for an LO. Of course, such a receiver has little tolerance of interferers.

REFERENCES

[1] B. Razavi et al., "Multiband UWB Transceivers," *Proc. CICC,* pp. 141–148, Sept 2005.
[2] B. Razavi, "Design Considerations for Direct-Conversion Receivers," *IEEE Trans. Circuits and Systems,* vol. 44, pp. 428–435, June 1997.
[3] A. A. Abidi, "Direct-conversion Radio Transceivers for Digital Communications," *IEEE Journal of Solid-State Circuits,* vol. 30, pp. 1399–1410, Dec. 1995.
[4] R. Hartley, "Modulation System," US Patent 1,666,206, April 1928.
[5] D. K. Weaver, "A Third Method of Generation and Detection of Single-Sideband Signals," *Proc. IRE,* vol. 44, pp. 1703–1705, Dec. 1956.
[6] J. Rudell et al., "A 1.9-GHz Wideband IF Double Conversion CMOS Receiver for Cordless Telephone Applications," *IEEE Journal of Solid-State Circuits,* vol. 32, pp. 2071–2088, Dec. 1997.
[7] L. Der and B. Razavi, "A 2-GHz CMOS Image-Reject Receiver with LMS Calibration," *IEEE Journal of Solid-State Circuits,* vol. 38, pp. 167–175, Feb. 2003.
[8] M. Gingell, "Single-Sideband Modulation Using Sequence Asymmetric Polyphase Networks," *Elec. Comm.*, vol. 48, pp. 21–25, 1973.
[9] S. Lerstaveesin and B. S. Song, "A Complex Image Rejection Circuit with Sign Detection Only," *IEEE J. Solid-State Circuits,* vol. 41, pp. 2693–2702, Dec. 2006.
[10] J. Crols and M. S. J. Steyaert, "A Single-Chip 900-MHz CMOS Receiver Front End with a High-Perfromance Low-IF Topology," *IEEE J. Solid-State Circuits,* vol. 30, pp. 1483–1492, Dec. 1995.
[11] F. Behbahani et al., "CMOS Mixers and Polyphase Filters for Large Image Rejection," *IEEE J. Solid-State Circuits,* vol. 36, pp. 873–887, June 2001.
[12] J. Crols and M. S. J. Steyaert, "Low-IF Topologies for High-Performance Analog Front Ends of Fully Integrated Receivers," *IEEE Tran. Circuits and Sys., II,* vol. 45, pp. 269–282, March 1998.
[13] K. Feher, *Wireless Digital Communications,* New Jersey: Prentice-Hall, 1995.
[14] R. Steele, Ed., *Mobile Radio Communications,* New Jersey: IEEE Press, 1992.
[15] B. Razavi, "A 900-MHz/1.8-GHz CMOS Transmitter for Dual-Band Applications," *IEEE Journal of Solid-State Circuits,* vol. 34, pp. 573–579, May 1999.

[16] R. Adler, "A Study of Locking Phenomena in Oscillators," *Proc. of the IEEE*, vol. 61, No. 10, pp. 1380–1385, Oct. 1973.
[17] B. Razavi, "A Study of Injection Locking and Pulling in Oscillators," *IEEE J. of Solid-State Circuits*, vol. 39, pp. 1415–1424, Sep. 2004.
[18] M. Zargari et al., "A 5-GHz CMOS Transceiver for IEEE 802.11a Wireless LAN Systems," *IEEE J. of Solid-State Circuits*, vol. 37, pp. 1688–1694, Dec. 2002.
[19] S. A. Sanielevici et al., "A 900-MHz Transceiver Chipset for Two-Way Paging Applications," *IEEE J. of Solid-State Circuits*, vol. 33, pp. 2160–2168, Dec. 1998.
[20] M. Conta, private communication, Feb. 2011.
[21] B. Razavi, "A 5.2-GHz CMOS Receiver with 62-dB Image Rejection," *IEEE Journal of Solid-State Circuits*, vol. 36, pp. 810–815, May 2001.
[22] A. Parsa and B. Razavi, "A New Transceiver Architecture for the 60-GHz Band," *IEEE Journal of Solid-State Circuits*, vol. 44, pp. 751–762, Mar. 2009.

PROBLEMS

4.1. For the sliding-IF architecture of Fig. 4.26(a), assume the ÷2 circuit is replaced with a ÷4 circuit.

   (a) Determine the required LO frequency range and steps.

   (b) Determine the image frequency range.

4.2. Since the image band of the sliding-IF receiver of Fig. 4.26(a) is narrower than the signal band, is it possible to design an 11g receiver whose image is confined to the GPS band? Explain your reasoning.

4.3. A sliding-IF receiver with $f_{LO} = (2/3)f_{in}$ is designed for the 11g band. Determine some of the mixing spurs that result from the harmonics of the first LO and the second LO. Assume the second IF is zero.

4.4. Consider the 11g sliding-IF receiver shown in Fig. 4.116.

   (a) Determine the required LO frequency range.

   (b) Determine the image frequency range.

   (c) Is this architecture preferable to that in Fig. 4.26(a)? Why?

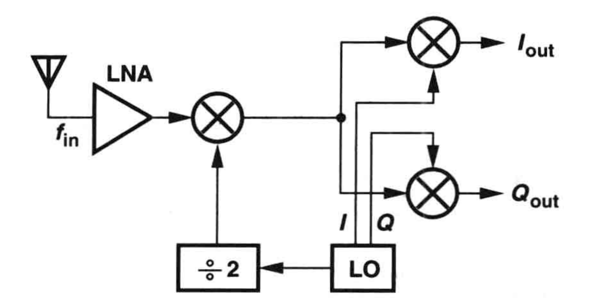


Figure 4.116 *Sliding-IF RX for 11g.*

4.5. Determine some of the mixing spurs in the architecture of Fig. 4.116.

4.6. The sliding-IF architecture shown in Fig. 4.117 is designed for the 11a band.

(a) Determine the image band.

(b) Determine the interferer frequencies that can appear in the output baseband as a result of mixing with the third harmonic of the first LO or the third harmonic of the second LO.

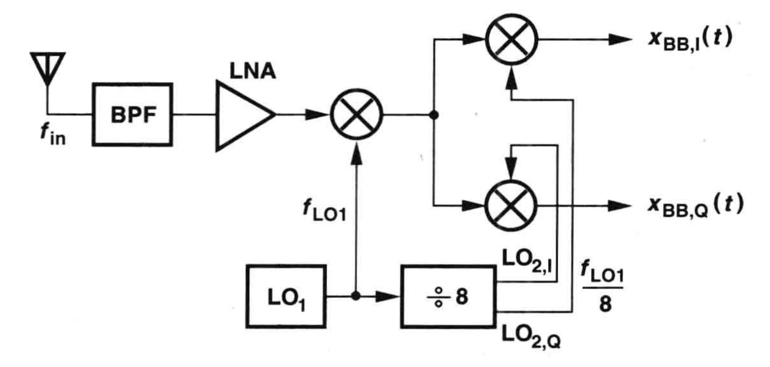


Figure 4.117 *Sliding-IF RX for 11a.*

4.7. Figure 4.118 shows a "half-RF" architecture, where $f_{LO} = f - in/2$ [21, 22].

(a) Assume the RF input is an asymmetrically-modulated signal. Sketch the spectra at the first and second IFs if the mixers are ideal multipliers.

(b) Repeat part (a) but assuming that the RF mixer also multiplies the RF signal by the third harmonic of the LO.

(c) The flicker noise of the LNA may be critical here. Explain why.

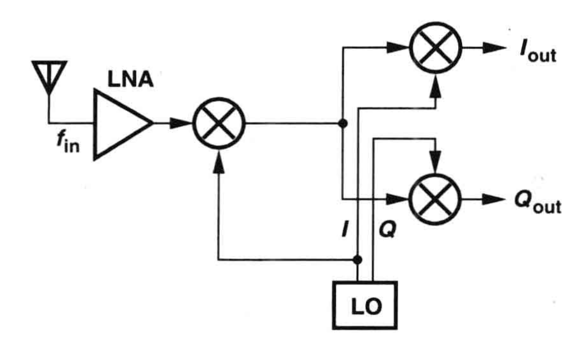


Figure 4.118 *Half-RF RX.*

4.8. Suppose an AM signal, $A(t)\cos\omega_c t$, is applied to a single mixer driven by an LO.

(a) If the LO waveform is given by $\cos\omega_c t$, determine the baseband signal.

(b) If the LO waveform is given by $\sin\omega_c t$, what happens? Why does this indicate the need for quadrature downconversion?

4.9. In this problem we wish to study an interesting effect that arises from LO leakage in direct-conversion receivers [20]. Consider the LO leakage in a direct-conversion receiver, $V_0 \cos \omega_{LO} t$. Suppose this leakage is added to an amplitude-modulated interferer, $V_{int}(t) \cos \omega_{int} t$, and the result experiences third-order nonlinearity in the LNA (or downconversion mixer).

(a) Determine the components near the carrier at the LNA output.

(b) Determine the resulting baseband components and whether they corrupt the desired signal.

4.10. In Example 4.24, how much gain must precede the given noise spectrum so that the penalty remains below 1 dB?

4.11. In Example 4.24, what flicker noise corner frequency is necessary if the penalty must remain below 1 dB?

4.12. An ASK waveform is applied to a direct-conversion receiver. Plot the baseband I and Q waveforms.

4.13. Does the quadrature mixing of Fig. 4.59(a) perform a Hilbert transform if the *upconverted* outputs at $\omega_c + \omega_{LO}$ are considered?

4.14. Does the Hartley architecture cancel the image if the IF low-pass filters are replaced with high-pass filters and the upconverted components are considered?

4.15. In the architecture of Fig. 4.64, assume the two resistors have a mismatch of ΔR. Compute the IRR.

4.16. Prove that the IRR of the Hartley architecture is given by $(\omega_{IF}/\Delta\omega)^2$ at an intermediate frequency of $\omega_{IF} + \Delta\omega$ if $\omega_{IF} = (R_1 C_1)^{-1}$.

4.17. Considering only the thermal noise of the resistors in Fig. 4.64 and assuming a voltage gain of A_1 for each mixer, determine the noise figure of the receiver with respect to a source impedance of R_D.

4.18. In the Weaver architecture of Fig. 4.67, both quadrature downconversions were performed with low-side injection. Study the other three combinations of high-side and low-side injection with the aid of signal spectra at nodes *A*-*F*.

4.19. Figure 4.119 shows three variants of the Hartley architecture. Explain which one(s) can reject the image.

4.20. If $\sin \omega_{LO} t$ and $\cos \omega_{LO} t$ in the Hartley architecture are swapped, does the RX still reject the image?

4.21. Repeat the above problem for the first or second LO in a Weaver architecture.

4.22. Using Eq. (4.96), compute the IRR of the receiver shown in Fig. 4.77(b) at an IF of $\omega + \Delta\omega$.

4.23. Assume a second-order parallel RLC tank is excited by a current source containing a component at ω_0 and another at $3\omega_0$. Prove that, if the tank resonates at $3\omega_0$, then the first harmonic is attenuated by approximately a factor of $8Q/3$ with respect to the third harmonic.

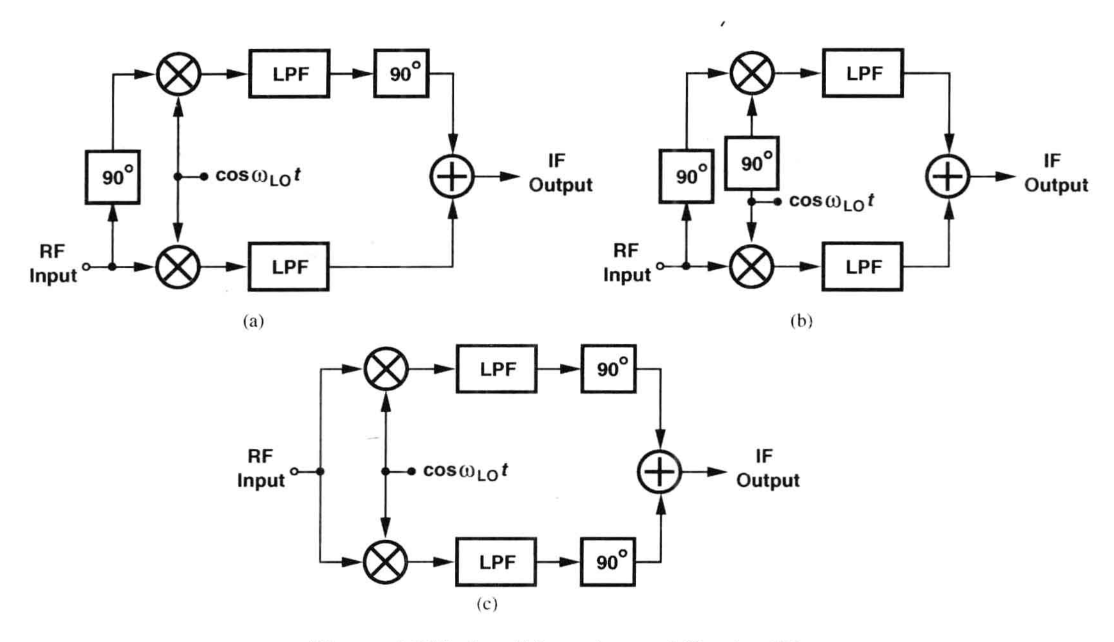


Figure 4.119 *Possible variants of Hartley RX.*

4.24. If the ÷2 circuit in Fig. 4.109 is replaced with a ÷4 circuit, study the spurious components arising from the third and fifth harmonics of the first and second LO frequencies.

4.25. The simplified Hartley architecture shown in Fig. 4.120 incorporates mixers having a voltage conversion gain of A_{mix} and an infinite input impedance. Taking into account only the noise of the two resistors, compute the noise figure of the receiver with respect to a source resistance of R_S at an IF of $1/(R_1C_1)$.

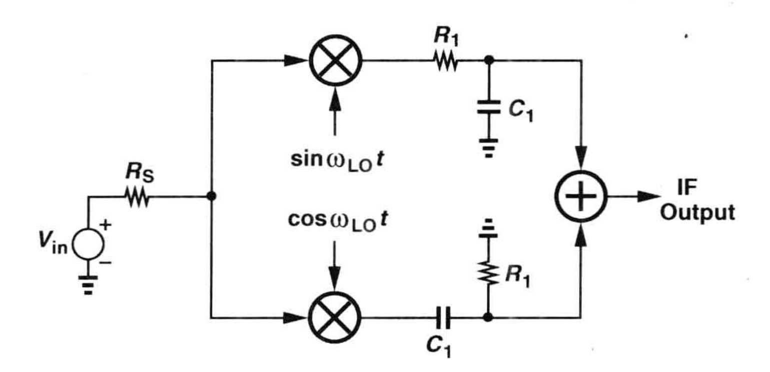


Figure 4.120 *Simplified Hartley RX.*

4.26. A dual-band receiver employing a Weaver architecture is shown in Fig. 4.121. The first LO frequency is chosen so as to create high-side injection for the 2.4-GHz band and low-side injection for the 5.2-GHz band. (The receiver operates only in one band at a given time.) Neglect the noise and nonlinearity of the receiver itself and assume

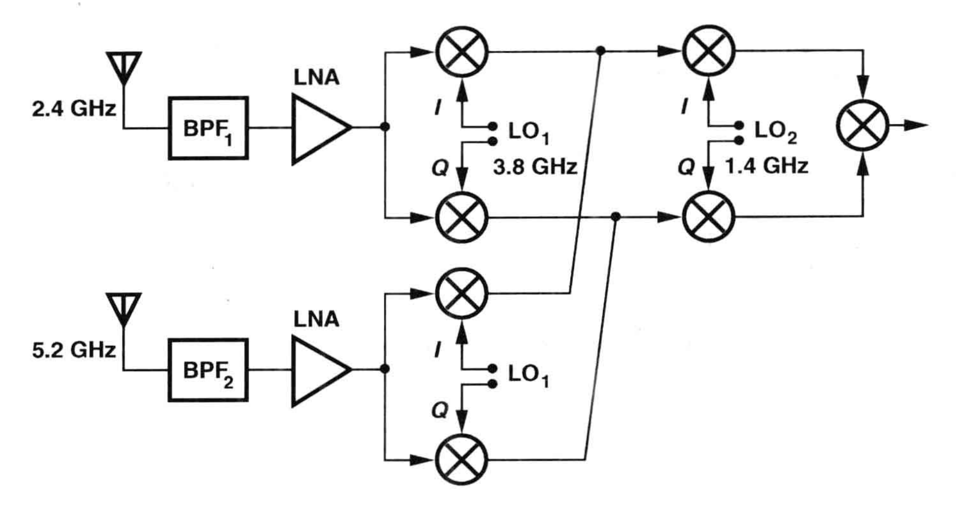


Figure 4.121 *Dual-band RX.*

an SNR of 20 dB is required for the signal to be detected properly. The Weaver architecture provides an image rejection ratio of 45 dB.

(a) Suppose the receiver must detect a −85-dBm signal in the 2.4-GHz mode while receiving at the same antenna a −10-dBm 5.2-GHz component as well. Determine the amount of rejection required of BPF_1 at 5.2 GHz.

(b) Suppose the receiver operates in the 5.2-GHz band but it also picks up a strong component at 7.2 GHz. It is possible for this component to be mixed with the third harmonics of LO_1 and LO_2 and appear in the baseband. Does the Weaver architecture prohibit this phenomenon? Explain in detail.

4.27. Consider the single-sideband mixer shown in Fig. 4.122. In the ideal case, the output has only one component at $\omega_1 + \omega_2$. Now suppose the ports sensing ω_2 suffer from third- and fifth-order nonlinearity. Plot the output spectrum if (a) $\omega_1 > 3\omega_2$ or (b) $\omega_1 < 3\omega_2$. Identify the frequency of each component.

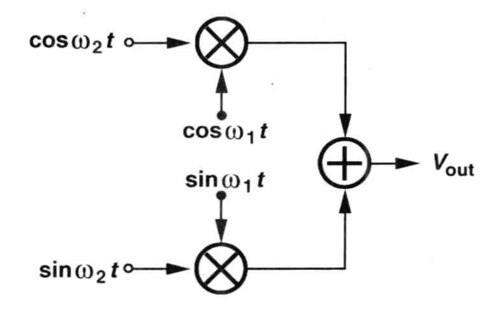
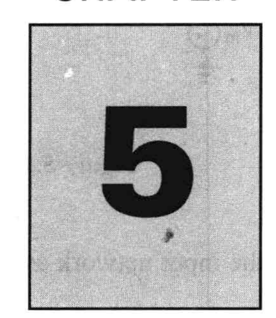


Figure 4.122 *SSB mixer.*

4.28. Explain why injection pulling is more serious in Fig. 4.114(b) than in Fig. 4.114(a).

CHAPTER



LOW-NOISE AMPLIFIERS

Following our system- and architecture-level studies in previous chapters, we move farther down to the circuit level in this and subsequent chapters. Beginning with the receive path, we describe the design of low-noise amplifiers. While our focus is on CMOS implementations, most of the concepts can be applied to other technologies as well. The outline of the chapter is shown below.

Basic LNA Topologies	Alternative LNA Topologies	Nonlinearity of LNAs
▪ CS Stage with Inductive Load ▪ CS Stage with Resistive Feedback ▪ CG Stage ▪ CS Stage with Inductive Degeneration	▪ Variants of CS LNA ▪ Noise-Cancelling LNAs ▪ Differential LNAs	▪ Nonlinearity Calculations ▪ Differential and Quasi-Differential LNAs

5.1 GENERAL CONSIDERATIONS

As the first active stage of receivers, LNAs play a critical role in the overall performance and their design is governed by the following parameters.

Noise Figure The noise figure of the LNA directly adds to that of the receiver. For a typical RX noise figure of 6 to 8 dB, it is expected that the antenna switch or duplexer contributes about 0.5 to 1.5 dB, the LNA about 2 to 3 dB, and the remainder of the chain about 2.5 to 3.5 dB. While these values provide a good starting point in the receiver design, the exact partitioning of the noise is flexible and depends on the performance of each stage in the chain. In modern RF electronics, we rarely design an LNA in isolation. Rather, we view and design the RF chain as one entity, performing many iterations among the stages.

To gain a better feel for a noise figure of 2 dB, consider the simple example in Fig. 5.1(a), where the noise of the LNA is represented by only a voltage source. Rearranging

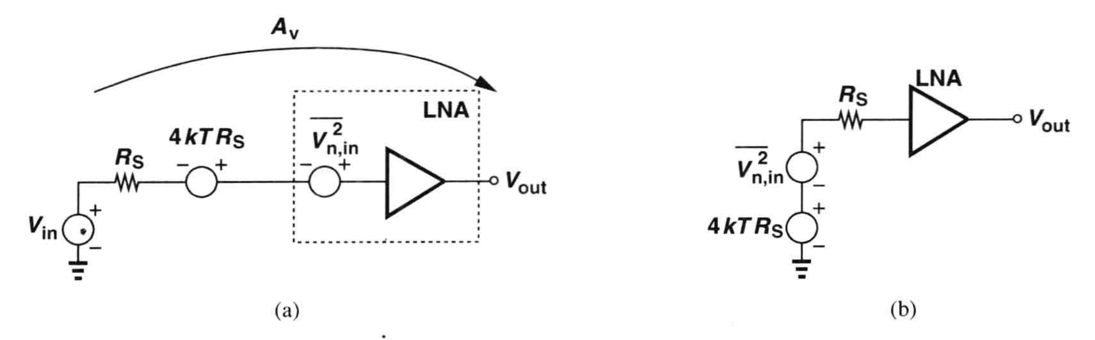


Figure 5.1 *(a) LNA with input-referred noise voltage, (b) simplified circuit.*

the input network as shown in Fig. 5.1(b), we have from Chapter 2

$$\text{NF} = \frac{\overline{V^2_{n,out}}}{A_v^2} \cdot \frac{1}{4kTR_S} \tag{5.1}$$

$$= 1 + \frac{\overline{V^2_{n,in}}}{4kTR_S}. \tag{5.2}$$

Thus, a noise figure of 2 dB with respect to a source impedance of 50 Ω translates to $\sqrt{\overline{V^2_{n,in}}} = 0.696\ \text{nV}/\sqrt{\text{Hz}}$, an extremely low value. For the gate-referred thermal noise voltage of a MOSFET, $4kT\gamma/g_m$, to reach this value, the g_m must be as high as $(29\ \Omega)^{-1}$ (if $\gamma = 1$). In this chapter, we assume $R_S = 50\ \Omega$.

Example 5.1

A student lays out an LNA and connects its input to a pad through a metal line 200 μm long. In order to minimize the input capacitance, the student chooses a width of 0.5 μm for the line. Assuming a noise figure of 2 dB for the LNA and a sheet resistance of 40 mΩ/□ for the metal line, determine the overall noise figure. Neglect the input-referred noise current of the LNA.

Solution:

We draw the equivalent circuit as shown in Fig. 5.2, pretending that the line resistance, R_L, is part of the LNA. The total input-referred noise voltage of the circuit inside the box is

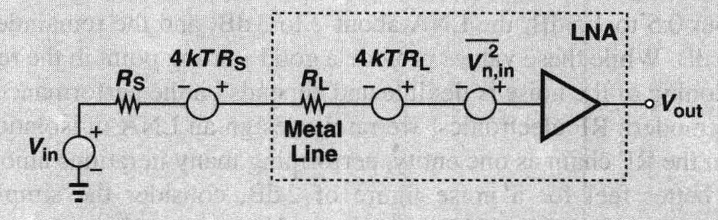


Figure 5.2 *LNA with metal resistance in series with its input.*

Example 5.1 (*Continued*)

therefore equal to $\overline{V_{n,in}^2} + 4kTR_L$. We thus write

$$NF_{tot} = 1 + \frac{\overline{V_{n,in}^2} + 4kTR_L}{4kTR_S} \tag{5.3}$$

$$= 1 + \frac{\overline{V_{n,in}^2}}{4kTR_S} + \frac{R_L}{R_S} \tag{5.4}$$

$$= NF_{LNA} + \frac{R_L}{R_S}, \tag{5.5}$$

where NF_{LNA} denotes the noise figure of the LNA without the line resistance. Since $NF_{LNA} = 2$ dB $\equiv 1.58$ and $R_L = (200/0.5) \times 40$ m$\Omega/\square = 16\ \Omega$, we have

$$NF_{tot} = 2.79 \text{ dB}. \tag{5.6}$$

The point here is that even small amounts of line or gate resistance can raise the noise figure of LNAs considerably.

The low noise required of LNAs limits the choice of the circuit topology. This often means that *only one transistor*—usually the input device—can be the dominant contributor to NF, thus ruling out configurations such as emitter or source followers.

Gain The gain of the LNA must be large enough to minimize the noise contribution of subsequent stages, specifically, the downconversion mixer(s). As described in Chapter 2, the choice of this gain leads to a compromise between the noise figure and the linearity of the receiver as a higher gain makes the nonlinearity of the subsequent stages more pronounced. In modern RF design, the LNA directly drives the downconversion mixer(s) with no impedance matching between the two. Thus, it is more meaningful and simpler to perform the chain calculations in terms of the voltage gain—rather than power gain—of the LNA.

It is important to note that the noise and IP_3 of the stage following the LNA are divided by *different* LNA gains. Consider the LNA/mixer cascade shown in Fig. 5.3(a), where the input-referred noise voltages are denoted by $\overline{V_{n,LNA}^2}$ and $\overline{V_{n,mixer}^2}$ and input noise currents

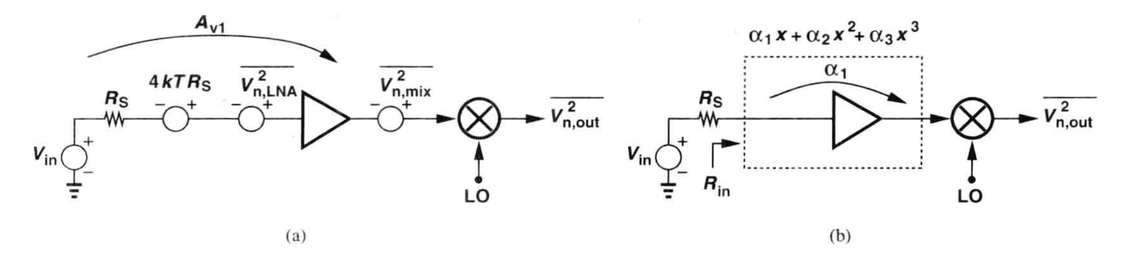


Figure 5.3 *Appropriate choice of gain for referring (a) noise and (b) IP_3 of a mixer to LNA input.*

are neglected. Assuming a unity voltage gain for the mixer for simplicity, we write the total output noise as $A_{v1}^2(\overline{V_{n,LNA}^2} + 4kTR_S) + \overline{V_{n,mix}^2}$. The overall noise figure is thus equal to

$$\text{NF}_{\text{tot}} = \frac{A_{v1}^2(\overline{V_{n,LNA}^2} + 4kTR_S) + \overline{V_{n,mix}^2}}{A_{v1}^2} \frac{1}{4kTR_S} \tag{5.7}$$

$$= \text{NF}_{\text{LNA}} + \frac{\overline{V_{n,mix}^2}}{A_{v1}^2} \cdot \frac{1}{4kTR_S}. \tag{5.8}$$

In other words, for NF calculations, the noise of the second stage is divided by the gain from the input voltage source to the LNA output.

Now consider the same cascade repeated in Fig. 5.3(b) with the nonlinearity of the LNA expressed as a third-order polynomial. From Chapter 2, we have[1]

$$\frac{1}{\text{IP}_{3,\text{tot}}^2} = \frac{1}{\text{IP}_{3,\text{LNA}}^2} + \frac{\alpha_1^2}{\text{IP}_{3,\text{mixer}}^2}. \tag{5.9}$$

In this case, α_1 denotes the voltage gain from the *input of the LNA* to its output. With input matching, we have $R_{in} = R_S$ and $\alpha_1 = 2A_{v1}$. That is, the mixer noise is divided by the *lower* gain and the mixer IP_3 by the *higher* gain—both against the designer's wish.

Input Return Loss The interface between the antenna and the LNA entails an interesting issue that divides analog designers and microwave engineers. Considering the LNA as a *voltage* amplifier, we may expect that its input impedance must ideally be infinite. From the noise point of view, we may precede the LNA with a transformation network to obtain minimum NF. From the signal *power* point of view, we may realize *conjugate* matching between the antenna and the LNA. Which one of these choices is preferable?

We make the following observations. (1) the (off-chip) band-select filter interposed between the antenna and the LNA is typically designed and characterized as a high-frequency device and with a standard termination of 50 Ω. If the load impedance seen by the filter (i.e., the LNA input impedance) deviates from 50 Ω significantly, then the passband and stopband characteristics of the filter may exhibit loss and ripple. (2) Even in the absence of such a filter, the antenna itself is designed for a certain real load impedance, suffering from uncharacterized loss if its load deviates from the desired real value or contains an imaginary component. Antenna/LNA co-design could improve the overall performance by allowing even non-conjugate matching, but it must be borne in mind that, if the antenna is shared with the transmitter, then its impedance must contain a negligible imaginary part so that it *radiates* the PA signal. (3) In practice, the antenna signal must travel a considerable distance on a printed-circuit board before reaching the receiver. Thus, poor matching at the RX input leads to significant reflections, an uncharacterized loss, and possibly voltage attenuation. For these reasons, the LNA is designed for a 50-Ω resistive input impedance. Since none of the above concerns apply to the other interfaces within the RX (e.g., between the LNA and the mixer or between the LO and the mixer), they are typically designed to maximize *voltage* swings rather than power transfer.

1. The IM_3 components arising from second-order terms are neglected.

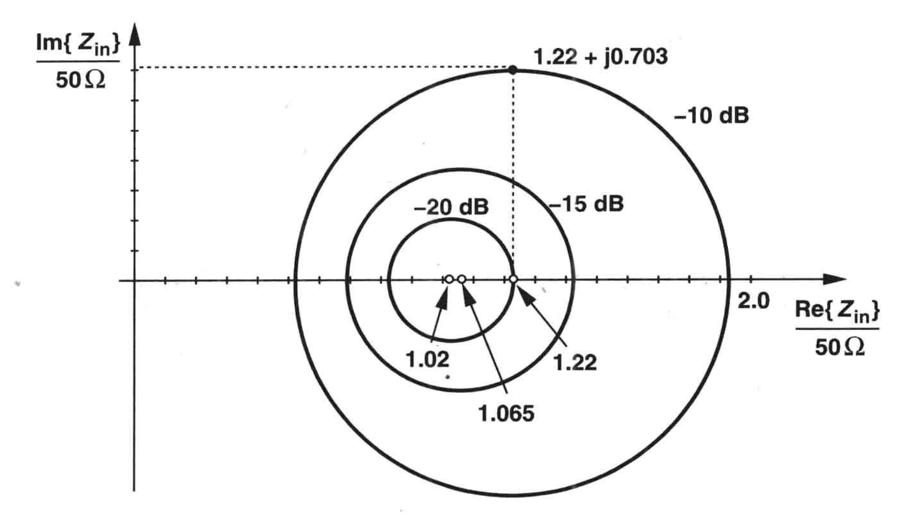


Figure 5.4 *Constant-Γ contours in the input impedance plane.*

The quality of the input match is expressed by the input "return loss," defined as the reflected power divided by the incident power. For a source impedance of R_S, the return loss is given by[2]

$$\Gamma = \left| \frac{Z_{in} - R_S}{Z_{in} + R_S} \right|^2, \tag{5.10}$$

where Z_{in} denotes the input impedance. An input return loss of -10 dB signifies that one-tenth of the power is reflected—a typically acceptable value. Figure 5.4 plots contours of constant Γ in the Z_{in} plane. Each contour is a circle with its center shown. For example, $Re\{Z_{in}\} = 1.22 \times 50\,\Omega = 61\,\Omega$ and $Im\{Z_{in}\} = 0.703 \times 50\,\Omega = 35.2\,\Omega$ yield $S_{11} = -10$ dB. In Problem 5.1, we derive the equations for these contours. We should remark that, in practice, a Γ of about -15 dB is targeted so as to allow margin for package parasitics, etc.

Stability Unlike the other circuits in a receiver, the LNA must interface with the "outside world," specifically, a poorly-controlled source impedance. For example, if the user of a cell phone wraps his/her hand around the antenna, the antenna impedance changes.[3] For this reason, the LNA must remain stable for all source impedances at *all frequencies*. One may think that the LNA must operate properly only in the frequency band of interest and not necessarily at other frequencies, but if the LNA begins to oscillate at any frequency, it becomes highly nonlinear and its gain is very heavily compressed.

A parameter often used to characterize the stability of circuits is the "Stern stability factor," defined as

$$K = \frac{1 + |\Delta|^2 - |S_{11}|^2 - |S_{22}|^2}{2|S_{21}||S_{12}|}, \tag{5.11}$$

2. Note that Γ is sometimes defined as $(Z_{in} - R_S)/(Z_{in} + R_S)$, in which case it is expressed in decibels by computing $20 \log \Gamma$ (rather than $10 \log \Gamma$).

3. In the presence of a front-end band-select filter, the LNA sees smaller changes in the source impedance.

where $\Delta = S_{11}S_{22} - S_{12}S_{21}$. If $K > 1$ and $\Delta < 1$, then the circuit is unconditionally stable, i.e., it does not oscillate with any combination of *source* and *load* impedances. In modern RF design, on the other hand, the load impedance of the LNA (the input impedance of the on-chip mixer) is relatively well-controlled, making K a pessimistic measure of stability. Also, since the LNA output is typically not matched to the input of the mixer, S_{22} is not a meaningful quantity in such an environment.

Example 5.2

A cascade stage exhibits a high reverse isolation, i.e., $S_{12} \approx 0$. If the output impedance is relatively high so that $S_{22} \approx 1$, determine the stability conditions.

Solution:

With $S_{12} \approx 0$ and $S_{22} \approx 1$,

$$K \approx \frac{1 - |S_{22}|^2}{2|S_{21}||S_{12}|} > 1 \tag{5.12}$$

and hence

$$|S_{21}| < \frac{1 - |S_{22}|^2}{2|S_{12}|}. \tag{5.13}$$

In other words, the forward gain must not exceed a certain value. For $\Delta < 1$, we have

$$S_{11} < 1, \tag{5.14}$$

concluding that the input *resistance* must remain positive.

The above example suggests that LNAs can be stabilized by maximizing their reverse isolation. As explained in Section 5.3, this point leads to two robust LNA topologies that are naturally stable and hence can be optimized for other aspects of their performance with no stability concerns. A high reverse isolation is also necessary for suppressing the LO leakage to the input of the LNA.

LNAs may become unstable due to ground and supply parasitic inductances resulting from the packaging (and, at frequencies of tens of gigahertz, the on-chip line inductances). For example, if the gate terminal of a common-gate transistor sees a large series inductance, the circuit may suffer from substantial feedback from the output to the input and become unstable at some frequency. For this reason, precautions in the design and layout as well as accurate package modeling are essential.

Linearity In most applications, the LNA does not limit the linearity of the receiver. Owing to the cumulative gain through the RX chain, the latter stages, e.g., the baseband amplifiers or filters tend to limit the overall input IP_3 or P_{1dB}. We therefore design and optimize LNAs with little concern for their linearity.

An exception to the above rule arises in "full-duplex" systems, i.e., applications that transmit and receive simultaneously (and hence incorporate FDD). Exemplified by the

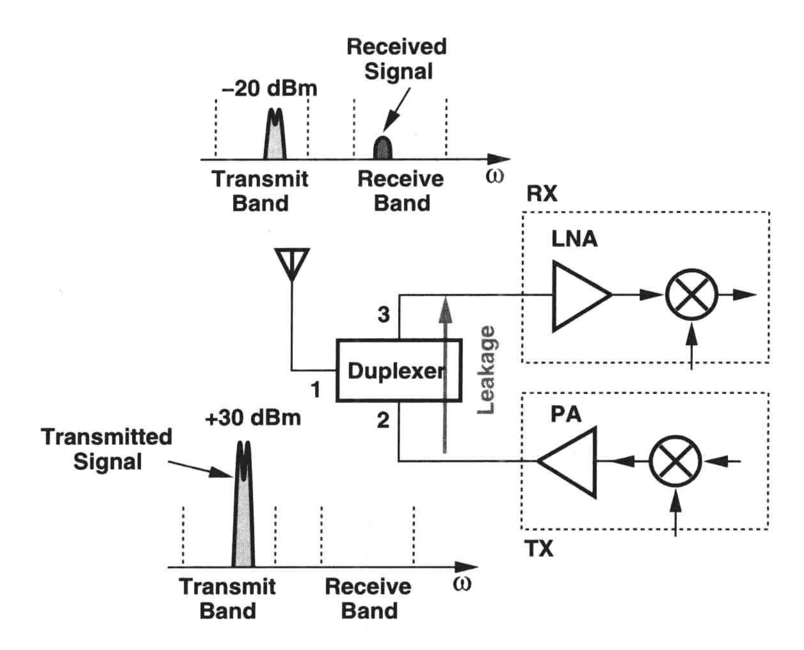


Figure 5.5 *TX leakage to RX in a full-duplex system.*

CDMA systems studied in Chapter 3, full-duplex operation must deal with the leakage of the strong transmitted signal to the receiver. To understand this issue, let us consider the front end shown in Fig. 5.5, where a duplexer separates the TX and RX bands. Modeling the duplexer as a three-port network, we note that S_{31} and S_{21} represent the losses in the RX and TX paths, respectively, and are about 1 to 2 dB. Unfortunately, leakages through the filter and the package yield a finite isolation between ports 2 and 3, as characterized by an S_{32} of about -50 dB. In other words, if the PA produces an average output power of $+30$ dBm (1 W), then the LNA experiences a signal level of -20 dBm in the TX band while sensing a much smaller received signal. Since the TX signal exhibits a variable envelope, its peak level may be about 2 dB higher. Thus, the receiver must remain uncompressed for an input level of -18 dBm. We must therefore choose a P_{1dB} of about -15 dBm to allow some margin.

Such a value for P_{1dB} may prove difficult to realize in a receiver. With an LNA gain of 15 to 20 dB, an input of -15 dBm yields an output of 0 to $+5$ dBm (632 to 1124 mV_{pp}), possibly compressing the LNA at *its output*. The LNA linearity is therefore critical. Similarly, the 1-dB compression point of the downconversion mixer(s) must reach 0 to $+5$ dBm. (The corresponding mixer IP_3 is roughly $+10$ to $+15$ dBm.) Thus, the mixer design also becomes challenging. For this reason, some CDMA receivers interpose an off-chip filter between the LNA and the mixer(s) so as to remove the TX leakage [1].

The linearity of the LNA also becomes critical in wideband receivers that may sense a large number of strong interferers. Examples include "ultra-wideband" (UBW), "software-defined," and "cognitive" radios.

Bandwidth The LNA must provide a relatively flat response for the frequency range of interest, preferably with less than 1 dB of gain variation. The LNA -3-dB bandwidth must therefore be substantially larger than the actual band so that the roll-off at the edges remains below 1 dB.

In order to quantify the difficulty in achieving the necessary bandwidth in a circuit, we often refer to its "fractional bandwidth," defined as the total −3-dB bandwidth divided by the center frequency of the band. For example, an 802.11g LNA requires a fractional bandwidth greater than 80 MHz/2.44 GHz = 0.0328.

Example 5.3

An 802.11a LNA must achieve a −3-dB bandwidth from 5 GHz to 6 GHz. If the LNA incorporates a second-order *LC* tank as its load, what is the maximum allowable tank *Q*?

Solution:

As illustrated in Fig. 5.6, the fractional bandwidth of an *LC* tank is equal to $\Delta\omega/\omega_0 = 1/Q$. Thus, the *Q* of the tank must remain less than 5.5 GHz/1 GHz = 5.5.

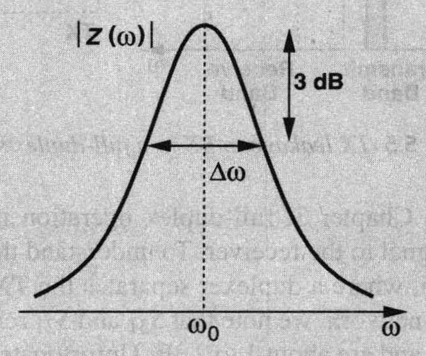


Figure 5.6 *Relationship between bandwidth and Q of a tank.*

LNA designs that must achieve a relatively large fractional bandwidth may employ a mechanism to *switch* the center frequency of operation. Depicted in Fig. 5.7(a) is an

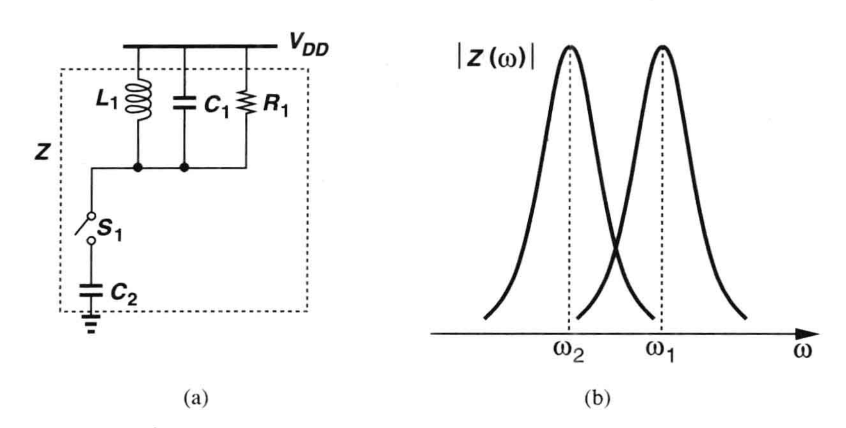


Figure 5.7 *(a) Band switching, (b) resulting frequency response.*

example, where an additional capacitor, C_2, can be switched into the tank, thereby changing the center frequency from $\omega_1 = 1/\sqrt{L_1C_1}$ to $\omega_2 = 1/\sqrt{L_1(C_1 + C_2)}$ [Fig. 5.7(b)]. We return to this concept in Section 5.5.

Power Dissipation The LNA typically exhibits a direct trade-off among noise, linearity, and power dissipation. Nonetheless, in most receiver designs, the LNA consumes only a small fraction of the overall power. In other words, the circuit's noise figure generally proves much more critical than its power dissipation.

5.2 PROBLEM OF INPUT MATCHING

As explained in Section 5.1, LNAs are typically designed to provide a 50-Ω input resistance and negligible input reactance. This requirement limits the choice of LNA topologies. In other words, we cannot begin with an arbitrary configuration, design it for a certain noise figure and gain, and then decide how to create input matching.

Let us first consider the simple common-source stage shown in Fig. 5.8, where C_F represents the gate-drain overlap capacitance. At very low frequencies, R_D is much smaller than the impedances of C_F and C_L and the input impedance is roughly equal to $[(C_{GS} + C_F)s]^{-1}$. At very high frequencies, C_F shorts the gate and drain terminals of M_1, yielding an input *resistance* equal to $R_D||(1/g_m)$. More generally, the reader can prove that the real and imaginary parts of the input admittance are, respectively, equal to

$$Re\{Y_{in}\} = R_D C_F \omega^2 \frac{C_F + g_m R_D (C_L + C_F)}{R_D^2 (C_L + C_F)^2 \omega^2 + 1} \tag{5.15}$$

$$Im\{Y_{in}\} = C_F \omega \frac{R_D^2 C_L (C_L + C_F)\omega^2 + 1 + g_m R_D}{R_D^2 (C_L + C_F)^2 \omega^2 + 1}. \tag{5.16}$$

Is it possible to select the circuit parameters so as to obtain $Re\{Y_{in}\} = 1/(50\,\Omega)$? For example, if $C_F = 10\,\text{fF}$, $C_L = 30\,\text{fF}$, $g_m R_D = 4$, and $R_D = 100\,\Omega$, then $Re\{Y_{in}\} = (7.8\,\text{k}\Omega)^{-1}$ at 5 GHz, far from $(50\,\Omega)^{-1}$. This is because C_F introduces little feedback at this frequency.

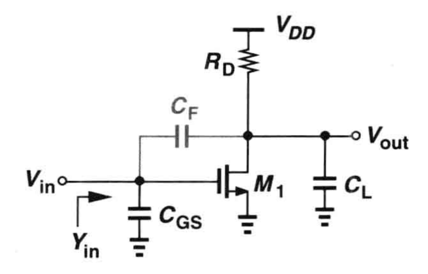


Figure 5.8 *Input admittance of a CS stage.*

Example 5.4

Why did we compute the input admittance rather than the input impedance for the circuit of Fig. 5.8.

Solution:

The choice of one over the other is somewhat arbitrary. In some circuits, it is simpler to compute Y_{in}. Also, if the input capacitance is cancelled by a *parallel* inductor, then $Im\{Y_{in}\}$ is more relevant. Similarly, a series inductor would cancel $Im\{Z_{in}\}$. We return to these concepts later in this chapter.

Can we employ simple resistive termination at the input? Illustrated in Fig. 5.9(a), such a topology is designed in three steps: (1) M_1 and R_D provide the required noise figure and gain, (2) R_P is placed in parallel with the input to provide $Re\{Z_{in}\} = 50\,\Omega$, and (3) an inductor is interposed between R_S and the input to cancel $Im\{Z_{in}\}$. Unfortunately, as explained in Chapter 2, the termination resistor itself yields a noise figure of $1 + R_S/R_P$. To calculate the noise figure at low frequencies, we can utilize Friis' equation[4] or simply treat the entire LNA as one circuit and, from Fig. 5.9(b), express the total output noise as

$$\overline{V_{n,out}^2} = 4kT(R_S||R_P)(g_mR_D)^2 + 4kT\gamma g_m R_D^2 + 4kTR_D, \tag{5.17}$$

where channel-length modulation is neglected. Since the voltage gain from V_{in} to V_{out} in Fig. 5.9(a) is equal to $-[R_P/(R_P + R_S)]g_mR_D$, the noise figure is given by

$$\text{NF} = 1 + \frac{R_S}{R_P} + \frac{\gamma R_S}{g_m(R_S||R_P)^2} + \frac{R_S}{g_m^2(R_S||R_P)^2 R_D}. \tag{5.18}$$

For $R_P \approx R_S$, the NF exceeds 3 dB—perhaps substantially.

The key point in the foregoing study is that the LNA must provide a 50-Ω input resistance *without* the thermal noise of a physical 50-Ω resistor. This becomes possible with the aid of active devices.

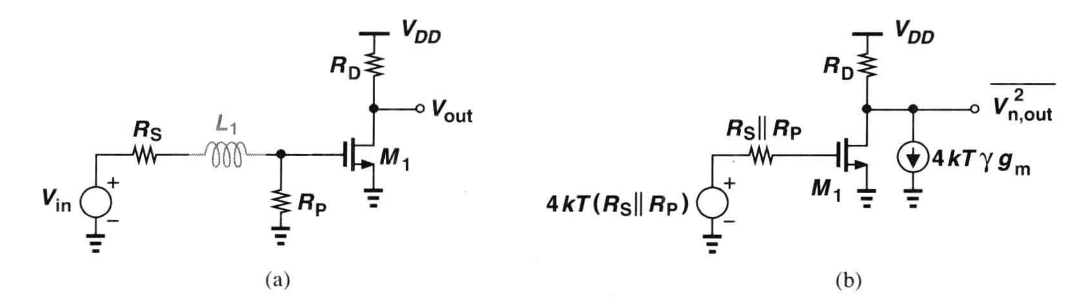


Figure 5.9 *(a) Use of resistive termination for matching, (b) simplified circuit.*

4. That is, consider R_P as one stage and the CS amplifier as another.

Example 5.5

A student decides to defy the above observation by choosing a *large* R_P and *transforming* its value down to R_S. The resulting circuit is shown in Fig. 5.10(a), where C_1 represents the input capacitance of M_1. (The input resistance of M_1 is neglected.) Can this topology achieve a noise figure less than 3 dB?

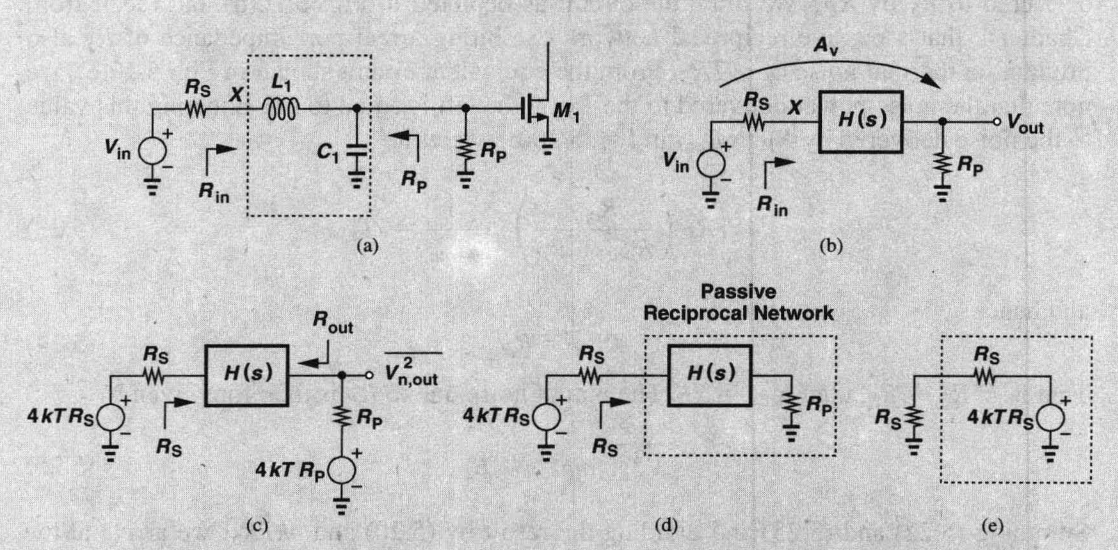


Figure 5.10 *(a) Use of matching circuit to transform the value of R_P, (b) general representation of (a), (c) inclusion of noise of R_P, (d) simplified circuit of (c), (e) simplified circuit of (d).*

Solution:

Consider the more general circuit in Fig. 5.10(b), where $H(s)$ represents a lossless network similar to L_1 and C_1 in Fig. 5.10(a). Since it is desired that $Z_{in} = R_S$, the power delivered by V_{in} to the input port of $H(s)$ is equal to $(V_{in,rms}/2)^2/R_S$. This power must also be delivered to R_P:

$$\frac{V^2_{in,rms}}{4R_S} = \frac{V^2_{out,rms}}{R_P}. \tag{5.19}$$

It follows that

$$|A_v|^2 = \frac{R_P}{4R_S}. \tag{5.20}$$

Let us now compute the output noise with the aid of Fig. 5.10(c). The output noise due to the noise of R_S is readily obtained from Eq. (5.19) by the substitution $\overline{V^2_{in,rms}} = 4kTR_S$:

$$\overline{V^2_{n,out}}|_{RS} = 4kTR_S \cdot \frac{R_P}{4R_S} \tag{5.21}$$

$$= kTR_P. \tag{5.22}$$

(Continues)

Example 5.5 *(Continued)*

But, how about the noise of R_P? We must first determine the value of R_{out}. To this end, we invoke the following thermodynamics principle: if R_S and R_P are in thermal equilibrium, then the noise power delivered by R_S to R_P must remain equal to the noise power delivered by R_P to R_S; otherwise, one heats up and the other cools down. How much is the noise delivered to R_S by R_P? We draw the circuit as depicted in Fig. 5.10(d) and recall from Chapter 2 that a passive reciprocal network exhibiting a real port impedance of R_S also produces a thermal noise of $4kTR_S$. From the equivalent circuit shown in Fig. 5.10(e), we note that the noise power delivered to the R_S on the left is equal to kT. Equating this value to the noise delivered by R_P to R_{out} in Fig. 5.10(c), we write

$$4kTR_P\left(\frac{R_{out}}{R_{out}+R_P}\right)^2\cdot\frac{1}{R_{out}}=kT \tag{5.23}$$

and hence

$$R_{out}=R_P. \tag{5.24}$$

That is, if $R_{in}=R_S$, then $R_{out}=R_P$. The output noise due to R_P is therefore given by

$$\overline{V_{n,out}^2}|_{RP}=kTR_P. \tag{5.25}$$

Summing (5.22) and (5.25) and dividing the result by (5.20) and $4kTR_S$, we arrive at the noise figure of the circuit (excluding M_1):

$$\text{NF}=2. \tag{5.26}$$

Unfortunately, the student has attempted to defy the laws of physics.

In summary, proper input (conjugate) matching of LNAs requires certain circuit techniques that yield a real part of 50 Ω in the input impedance without the noise of a 50-Ω resistor. We study such techniques in the next section.

5.3 LNA TOPOLOGIES

Our preliminary studies thus far suggest that the noise figure, input matching, and gain constitute the principal targets in LNA design. In this section, we present a number of LNA topologies and analyze their behavior with respect to these targets. Table 5.1 provides an overview of these topologies.

5.3.1 Common-Source Stage with Inductive Load

As noted in Section 5.1, a CS stage with resistive load (Fig. 5.8) proves inadequate because it does not provide proper matching. Furthermore, the output node time constant may prohibit operation at high frequencies. In general, the trade-off between the voltage gain and

Table 5.1 *Overview of LNA topologies.*

Common–Source Stage with	Common–Gate Stage with	Broadband Topologies
■ Inductive Load ■ Resistive Feedback ■ Cascode, Inductive Load, Inductive Degeneration	■ Inductive Load ■ Feedback ■ Feedforward ■ Cascode and Inductive Load	■ Noise–Cancelling LNAs ■ Reactance–Cancelling LNAs

the supply voltage in this circuit makes it less attractive as the latter scales down with technology. For example, at low frequencies,

$$|A_v| = g_m R_D \tag{5.27}$$

$$= \frac{2I_D}{V_{GS} - V_{TH}} \cdot \frac{V_{RD}}{I_D} \tag{5.28}$$

$$= \frac{2V_{RD}}{V_{GS} - V_{TH}}, \tag{5.29}$$

where V_{RD} denotes the dc voltage drop across R_D and is limited by V_{DD}. With channel-length modulation, the gain is even lower.

In order to circumvent the trade-off expressed by Eq. (5.29) and also operate at higher frequencies, the CS stage can incorporate an inductive load. Illustrated in Fig. 5.11(a), such a topology operates with very low supply voltages because the inductor sustains a smaller dc voltage drop than a resistor does. (For an ideal inductor, the dc drop is zero.) Moreover, L_1 resonates with the total capacitance at the output node, affording a much higher operation frequency than does the resistively-loaded counterpart of Fig. 5.8.

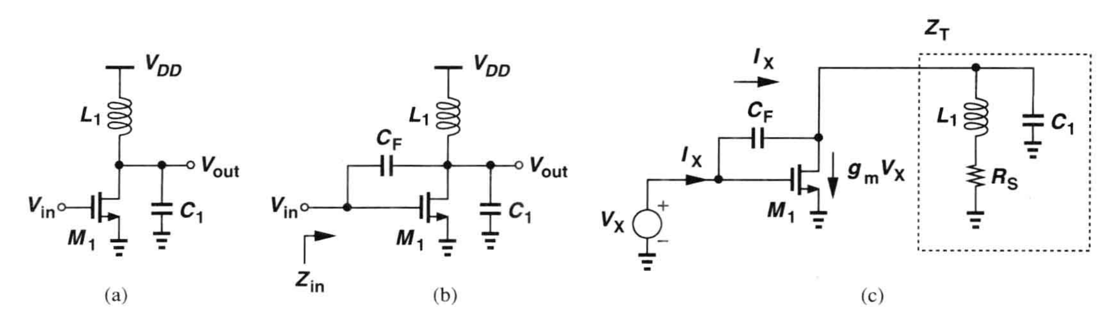


Figure 5.11 *(a) Inductively-loaded CS stage, (b) input impedance in the presence of C_F, (c) equivalent circuit.*

How about the input matching? We consider the more complete circuit shown in Fig. 5.11(b), where C_F denotes the gate-drain overlap capacitance. Ignoring the gate-source capacitance of M_1 for now, we wish to compute Z_{in}. We redraw the circuit as depicted in Fig. 5.11(c) and note that the current flowing through the output parallel tank is equal to

$I_X - g_m V_X$. In this case, the inductor loss is modeled by a *series* resistance, R_S, because this resistance varies much less with frequency than the equivalent parallel resistance does.[5] The tank impedance is given by

$$Z_T = \frac{L_1 s + R_S}{L_1 C_1 s^2 + R_S C_1 s + 1}, \tag{5.30}$$

and the tank voltage by $(I_X - g_m V_X)Z_T$. Adding the voltage drop across C_F to the tank voltage, we have

$$V_X = \frac{I_X}{C_F s} + (I_X - g_m V_X)Z_T. \tag{5.31}$$

Substitution of Z_T from (5.30) gives

$$Z_{in}(s) = \frac{V_X}{I_X} = \frac{L_1(C_1 + C_F)s^2 + R_S(C_1 + C_F)s + 1}{[L_1 C_1 s^2 + (R_S C_1 + g_m L_1)s + 1 + g_m R_S]C_F s}. \tag{5.32}$$

For $s = j\omega$,

$$Z_{in}(j\omega) = \frac{1 - L_1(C_1 + C_F)\omega^2 + jR_S(C_1 + C_F)\omega}{[-(R_S C_1 + g_m L_1)\omega + j(g_m R_S - L_1 C_1 \omega^2 + 1)]C_F \omega}. \tag{5.33}$$

Since the real part of a complex fraction $(a + jb)/(c + jd)$ is equal to $(ac + bd)/(c^2 + d^2)$, we have

$$Re\{Z_{in}\} = \frac{[1 - L_1(C_1 + C_F)\omega^2][-(R_S C_1 + g_m L_1)\omega] + R_S(C_1 + C_F)(g_m R_S - L_1 C_1 \omega^2 + 1)\omega^2}{D}, \tag{5.34}$$

where D is a positive quantity. It is thus possible to select the values so as to obtain $Re\{Z_{in}\} = 50\ \Omega$.

While providing the possibility of $Re\{Z_{in}\} = 50\ \Omega$ at the frequency of interest, the feedback capacitance in Fig. 5.11(b) gives rise to a *negative* input resistance at other frequencies, potentially causing instability. To investigate this point, let us rewrite Eq. (5.34) as

$$Re\{Z_{in}\} = \frac{g_m L_1^2(C_1 + C_F)\omega^2 + R_S(1 + g_m R_S)(C_1 + C_F) - (R_S C_1 + g_m L_1)}{D}\omega. \tag{5.35}$$

We note that the numerator falls to zero at a frequency given by

$$\omega_1^2 = \frac{R_S C_1 + g_m L_1 - (1 + g_m R_S)R_S(C_1 + C_F)}{g_m L_1^2(C_1 + C_F)}. \tag{5.36}$$

5. For example, if R_S simply represents the low-frequency resistance of the wire, its value remains constant and $Q = L\omega/R_S$ rises linearly with frequency. For a parallel resistance, R_P, to allow such a behavior for $Q = R_P/(L\omega)$, the resistance must rise in proprotion to ω^2 rather than remain constant.

Thus, at this frequency (if it exists), $Re\{Z_{in}\}$ changes sign. For example, if $C_F = 10\,\text{fF}$, $C_1 = 30\,\text{fF}$, $g_m = (20\,\Omega)^{-1}$, $L_1 = 5$ nH, and $R_S = 20\,\Omega$, then $g_m L_1$ dominates in the numerator, yielding $\omega_1^2 \approx [L_1(C_1 + C_F)]^{-1}$ and hence $\omega_1 \approx 2\pi \times (11.3\,\text{GHz})$.

It is possible to "neutralize" the effect of C_F in some frequency range through the use of parallel resonance (Fig. 5.12), but, since C_F is relatively small, L_F must assume a large value, thereby introducing significant parasitic capacitances at the input and output (and even *between* the input and output) and degrading the performance. For these reasons, this topology is rarely used in modern RF design.

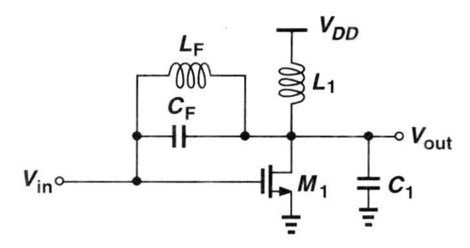


Figure 5.12 *Neutralization of C_F by L_F.*

5.3.2 Common-Source Stage with Resistive Feedback

If the frequency of operation remains an order of magnitude lower than the f_T of the transistor, the feedback CS stage depicted in Fig. 5.13(a) may be considered as a possible candidate. Here, M_2 operates as a current source and R_F senses the output voltage and returns a current to the input. We wish to design this stage for an input resistance equal to R_S and a relatively low noise figure.

If channel-length modulation is neglected, we have from Fig. 5.13(b),

$$R_{in} = \frac{1}{g_{m1}} \tag{5.37}$$

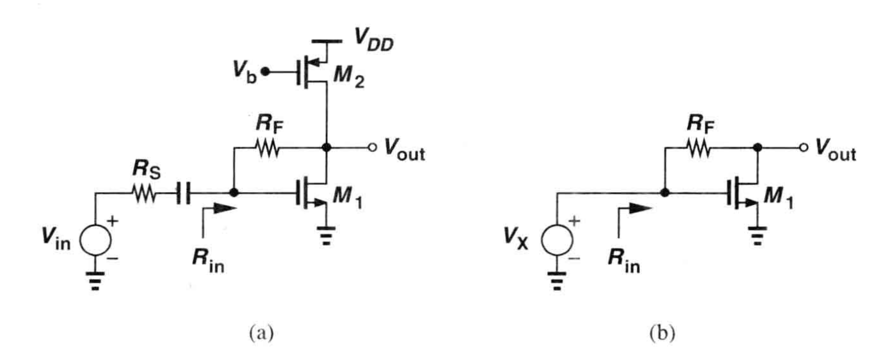


Figure 5.13 *(a) CS stage with resistive feedback, (b) simplified circuit.*

because R_F is simply in series with an ideal current source and M_1 appears as a diode-connected device. We must therefore choose

$$g_{m1} = \frac{1}{R_S}. \tag{5.38}$$

Figure 5.13(b) also implies that the small-signal drain current of M_1, $g_{m1}V_X$, entirely flows through R_F, generating a voltage drop of $g_{m1}V_XR_F$. It follows that

$$V_X - g_{m1}V_XR_F = V_{out} \tag{5.39}$$

and hence

$$\frac{V_{out}}{V_X} = 1 - g_{m1}R_F \tag{5.40}$$

$$= 1 - \frac{R_F}{R_S}. \tag{5.41}$$

In practice, $R_F \gg R_S$, and the voltage gain from V_{in} to V_{out} in Fig. 5.13(a) is equal to

$$A_v = \frac{1}{2}\left(1 - \frac{R_F}{R_S}\right) \tag{5.42}$$

$$\approx -\frac{R_F}{R_S}. \tag{5.43}$$

In contrast to the resistively-loaded CS stage of Fig. 5.8, this circuit does not suffer from a direct trade-off between gain and supply voltage because R_F carries no bias current.

Let us determine the noise figure of the circuit, assuming that $g_{m1} = 1/R_S$. We first compute the noise contributions of R_F, M_1, and M_2 at the output. From Fig. 5.14(a), the reader can show the noise of R_F appears at the output in its entirety:

$$\overline{V^2_{n,out}}|_{RF} = 4kTR_F. \tag{5.44}$$

The noise currents of M_1 and M_2 flow through the output impedance of the circuit, R_{out}, as shown in Fig. 5.14(b). The reader can prove that

$$R_{out} = \left[\frac{1}{g_{m1}}\left(1 + \frac{R_F}{R_S}\right)\right] || (R_F + R_S) \tag{5.45}$$

$$= \frac{1}{2}(R_F + R_S). \tag{5.46}$$

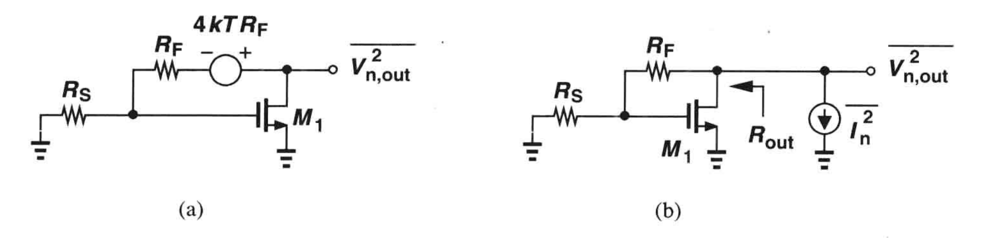


Figure 5.14 *Effect of noise of (a) R_F and (b) M_1 in CS stage.*

It follows that

$$\overline{V_{n,out}^2}|_{M1,M2} = 4kT\gamma(g_{m1} + g_{m2})\frac{(R_F + R_S)^2}{4}. \tag{5.47}$$

The noise of R_S is multiplied by the gain when referred to the output, and the result is divided by the gain when referred to the input. We thus have

$$\text{NF} = 1 + \frac{4R_F}{R_S\left(1 - \frac{R_F}{R_S}\right)^2} + \frac{\gamma(g_{m1} + g_{m2})(R_F + R_S)^2}{\left(1 - \frac{R_F}{R_S}\right)^2 R_S} \tag{5.48}$$

$$\approx 1 + \frac{4R_S}{R_F} + \gamma(g_{m1} + g_{m2})R_S \tag{5.49}$$

$$\approx 1 + \frac{4R_S}{R_F} + \gamma + \gamma g_{m2}R_S. \tag{5.50}$$

For $\gamma \approx 1$, the NF exceeds 3 dB even if $4R_S/R_F + \gamma g_{m2}R_S \ll 1$.

Example 5.6

Express the fourth term on the right-hand side of Eq. (5.50) in terms of transistor overdrive voltages.

Solution:

Since $g_m = 2I_D/(V_{GS} - V_{TH})$, we write $g_{m2}R_S = g_{m2}/g_{m1}$ and

$$\frac{g_{m2}}{g_{m1}} = \frac{(V_{GS} - V_{TH})_1}{|V_{GS} - V_{TH}|_2}. \tag{5.51}$$

That is, the fourth term becomes negligible only if the overdrive of the current source remains much higher than that of M_1—a difficult condition to meet at low supply voltages because $|V_{DS2}| = V_{DD} - V_{GS1}$. We should also remark that heavily velocity-saturated MOSFETs have a transconductance given by $g_m = I_D/(V_{GS} - V_{TH})$ and still satisfy (5.51).

Example 5.7

In the circuit of Fig. 5.15, the PMOS current source is converted to an "active load," amplifying the input signal. The idea is that, if M_2 amplifies the input in addition to injecting noise to the output, then the noise figure may be lower. Neglecting channel-length modulation, calculate the noise figure. (Current source I_1 defines the bias current, and C_1 establishes an ac ground at the source of M_2.)

(Continues)

Example 5.7 *(Continued)*

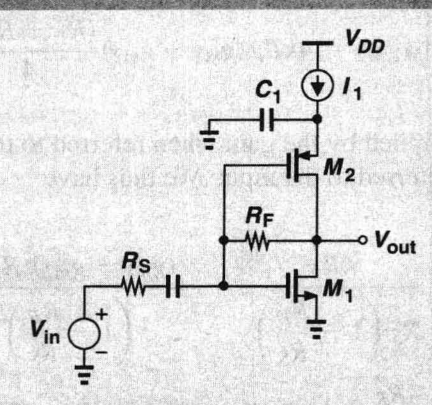


Figure 5.15 *CS stage with active load.*

Solution:

For small-signal operation, M_1 and M_2 appear *in parallel*, behaving as a single transistor with a transconductance of $g_{m1} + g_{m2}$. Thus, for input matching, $g_{m1} + g_{m2} = 1/R_S$. The noise figure is still given by Eq. (5.49), except that $\gamma(g_{m1} + g_{m2})R_S = \gamma$. That is,

$$\mathrm{NF} \approx 1 + \frac{4R_S}{R_F} + \gamma. \tag{5.52}$$

This circuit is therefore superior, but it requires a supply voltage equal to $V_{GS1} + |V_{GS2}| + V_{I1}$, where V_{I1} denotes the voltage headroom necessary for I_1.

5.3.3 Common-Gate Stage

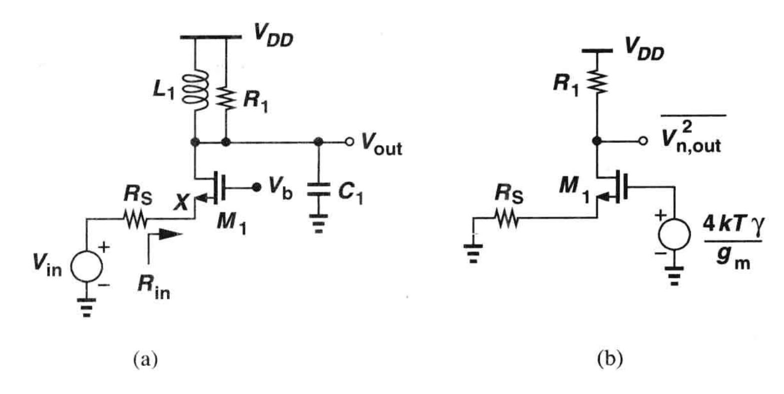
The low input impedance of the common-gate (CG) stage makes it attractive for LNA design. Since a resistively-loaded stage suffers from the same gain-headroom trade-off as its CS counterpart, we consider only a CG circuit with inductive loading [Fig. 5.16(a)]. Here, L_1 resonates with the total capacitance at the output node (including the input capacitance of the following stage), and R_1 represents the loss of L_1. If channel-length modulation and body effect are neglected, $R_{in} = 1/g_m$. Thus, the dimensions and bias current of M_1 are chosen so as to yield $g_m = 1/R_S = (50\,\Omega)^{-1}$. The voltage gain from X to the output node at the output resonance frequency is then equal to

$$\frac{V_{out}}{V_X} = g_m R_1 \tag{5.53}$$

$$= \frac{R_1}{R_S} \tag{5.54}$$

and hence $V_{out}/V_{in} = R_1/(2R_S)$.

Let us now determine the noise figure of the circuit under the condition $g_m = 1/R_S$ and at the resonance frequency. Modeling the thermal noise of M_1 as a voltage source in series



(a) (b)

Figure 5.16 *(a) CG stage, (b) effect of noise of M_1.*

with its gate, $\overline{V_{n1}^2} = 4kT\gamma/g_m$ [Fig. 5.16(b)], and multiplying it by the gain from the *gate* of M_1 to the output, we have

$$\overline{V^2_{n,out}}|_{M1} = \frac{4kT\gamma}{g_m}\left(\frac{R_1}{R_S + \frac{1}{g_m}}\right)^2 \tag{5.55}$$

$$= kT\gamma\frac{R_1^2}{R_S}. \tag{5.56}$$

The output noise due to R_1 is simply equal to $4kTR_1$. To obtain the noise figure, we divide the output noise due to M_1 and R_1 by the gain and $4kTR_S$ and add unity to the result:

$$\text{NF} = 1 + \frac{\gamma}{g_m R_S} + \frac{R_S}{R_1}\left(1 + \frac{1}{g_m R_S}\right)^2 \tag{5.57}$$

$$= 1 + \gamma + 4\frac{R_S}{R_1}. \tag{5.58}$$

Even if $4R_S/R_1 \ll 1 + \gamma$, the NF still reaches 3 dB (with $\gamma \approx 1$), a price paid for the condition $g_m = 1/R_S$. In other words, a higher g_m yields a lower NF but also a lower input resistance. In Problem 5.8, we show that the NF can be lower if some impedance mismatch is permitted at the input.

Example 5.8

We wish to provide the bias current of the CG stage by a current source or a resistor (Fig. 5.17). Compare the additional noise in these two cases.

(Continues)

Example 5.8 (*Continued*)

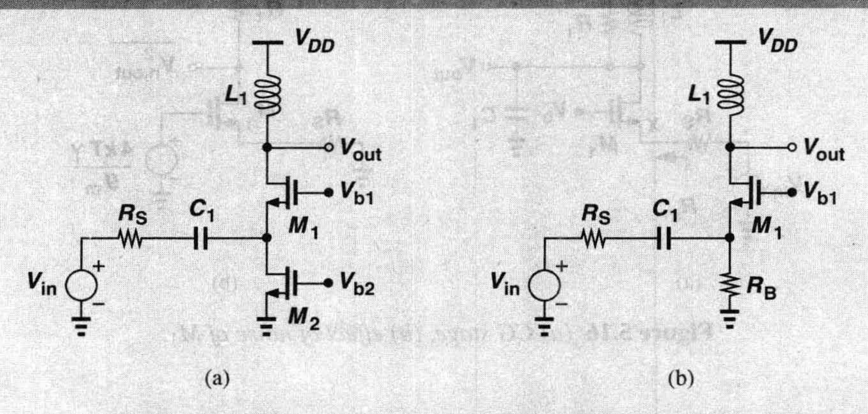


Figure 5.17 *CG stage biasing with (a) current source and (b) resistor.*

Solution:

For a given V_{b1} and V_{GS1}, the source voltages of M_1 in the two cases are equal and hence V_{DS2} is equal to the voltage drop across R_B ($=V_{RB}$). Operating in saturation, M_2 requires that $V_{DS2} \geq V_{GS2} - V_{TH2}$. We express the noise current of M_2 as

$$\overline{I^2_{n,M2}} = 4kT\gamma g_{m2} \tag{5.59}$$

$$= 4kT\gamma \frac{2I_D}{V_{GS2} - V_{TH2}}, \tag{5.60}$$

and that of R_B as

$$\overline{I^2_{n,RB}} = \frac{4kT}{R_B} \tag{5.61}$$

$$= 4kT\frac{I_D}{V_{RB}}. \tag{5.62}$$

Since $V_{GS2} - V_{TH2} \leq V_{RB}$, the noise contribution of M_2 is about twice that of R_B (for $\gamma \approx 1$). Additionally, M_2 may introduce significant capacitance at the input node.

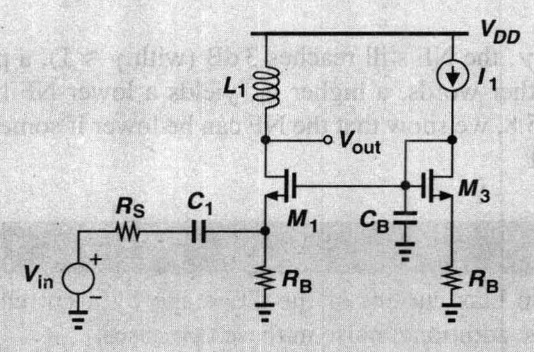


Figure 5.18 *Proper biasing of CG stage.*

Example 5.8 *(Continued)*

The use of a resistor is therefore preferable, so long as R_B is much greater than R_S so that it does not attenuate the input signal. Note that the input capacitance due to M_1 may still be significant. We will return to this issue later. Figure 5.18 shows an example of proper biasing in this case.

In deep-submicron CMOS technologies, channel-length modulation significantly impacts the behavior of the CG stage. As shown in Fig. 5.19, the positive feedback through r_O *raises* the input impedance. Since the drain-source current of M_1 (without r_O) is equal to $-g_m V_X$ (if body effect is neglected), the current flowing through r_O is given by $I_X - g_m V_X$, yielding a voltage drop of $r_O(I_X - g_m V_X)$ across it. Also, I_X flows through the output tank, producing a voltage of $I_X R_1$ at the resonance frequency. Adding this voltage to the drop across r_O and equating the result to V_X, we obtain

$$V_X = r_O(I_X - g_m V_X) + I_X R_1. \tag{5.63}$$

That is,

$$\frac{V_X}{I_X} = \frac{R_1 + r_O}{1 + g_m r_O}. \tag{5.64}$$

If the intrinsic gain, $g_m r_O$, is much greater than unity, then $V_X/I_X \approx 1/g_m + R_1/(g_m r_O)$. However, in today's technology, $g_m r_O$ hardly exceeds 10. Thus, the term $R_1/(g_m r_O)$ may become comparable with or even *exceed* the term $1/g_m$, yielding an input resistance substantially higher than 50 Ω.

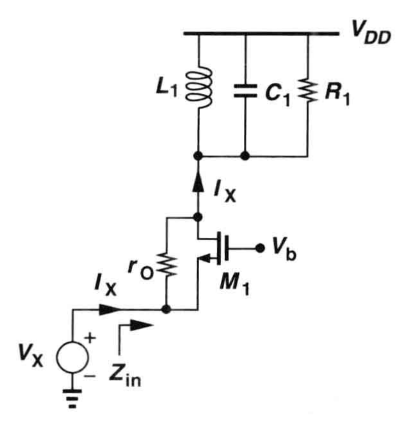


Figure 5.19 *Input impedance of CG stage in the presence of r_O.*

Example 5.9

Neglecting the capacitances of M_1 in Fig. 5.19, plot the input impedance as a function of frequency.

(Continues)

Example 5.9 *(Continued)*

Solution:

At very low or very high frequencies, the tank assumes a low impedance, yielding $R_{in} = 1/g_m$ [or $1/(g_m + g_{mb})$ if body effect is considered]. Figure 5.20 depicts the behavior.

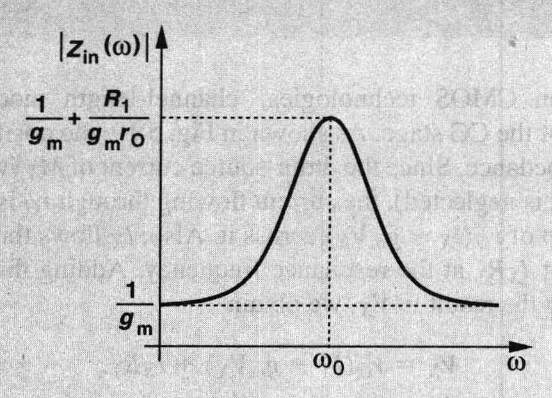


Figure 5.20 *Input impedance of CG stage with a resonant load.*

With the strong effect of R_1 on R_{in}, we must equate the actual input resistance to R_S to guarantee input matching:

$$R_S = \frac{R_1 + r_O}{1 + g_m r_O}. \tag{5.65}$$

The reader can prove that the voltage gain of the CG stage shown in Fig. 5.16(a) with a finite r_O is expressed as

$$\frac{V_{out}}{V_{in}} = \frac{g_m r_O + 1}{r_O + g_m r_O R_S + R_S + R_1} R_1, \tag{5.66}$$

which, from Eq. (5.65), reduces to

$$\frac{V_{out}}{V_{in}} = \frac{g_m r_O + 1}{2\left(1 + \dfrac{r_O}{R_1}\right)}. \tag{5.67}$$

This is a disturbing result! If r_O and R_1 are comparable, then the voltage gain is on the order of $g_m r_O/4$, a very low value.

In summary, the input impedance of the CG stage is too low if channel-length modulation is neglected and too high if it is not! A number of circuit techniques have been introduced to deal with the former case (Section 5.3.5), but in today's technology, we face the latter case.

In order to alleviate the above issue, the channel length of the transistor can be increased, thus reducing channel-length modulation and raising the achievable $g_m r_O$. Since the device width must also increase proportionally so as to retain the transconductance value, the gate-source capacitance of the transistor rises considerably, degrading the input return loss.

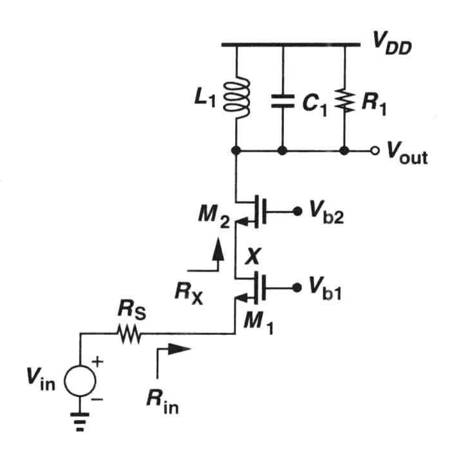


Figure 5.21 *Cascode CG stage.*

Cascode CG Stage An alternative approach to lowering the input impedance is to incorporate a cascode device as shown in Fig. 5.21. Here, the resistance seen looking into the source of M_2 is given by Eq. (5.64):

$$R_X = \frac{R_1 + r_{O2}}{1 + g_{m2}r_{O2}}. \tag{5.68}$$

This load resistance is now transformed to a lower value by M_1, again according to (5.64):

$$R_{in} = \left(\frac{R_1 + r_{O1}}{1 + g_{m2}r_{O2}} + r_{O1}\right) \div (1 + g_{m1}r_{O1}). \tag{5.69}$$

If $g_m r_O \gg 1$, then

$$R_{in} \approx \frac{1}{g_{m1}} + \frac{R_1}{g_{m1}r_{O1}g_{m2}r_{O2}} + \frac{1}{g_{m1}r_{O1}g_{m2}}. \tag{5.70}$$

Since R_1 is divided by the product of two intrinsic gains, its effect remains negligible. Similarly, the third term is much less than the first if g_{m1} and g_{m2} are roughly equal. Thus, $R_{in} \approx 1/g_{m1}$.

The addition of the cascode device entails two issues: the noise contribution of M_2 and the voltage headroom limitation due to stacking two transistors. To quantify the former, we consider the equivalent circuit shown in Fig. 5.22(a), where R_S ($= 1/g_{m1}$) and M_1 are replaced with an output resistance equal to $2r_{O1}$ (why?), and $C_X = C_{DB1} + C_{GD1} + C_{SB2}$. For simplicity, we have also replaced the tank with a resistor R_1, i.e., the output node has a broad bandwidth. Neglecting the gate-source capacitance, channel-length modulation, and body effect of M_2, we express the transfer function from V_{n2} to the output at the resonance frequency as

$$\frac{V_{n,out}}{V_{n2}}(s) = \frac{R_1}{\dfrac{1}{g_{m2}} + (2r_{O1})||\dfrac{1}{C_X s}} \tag{5.71}$$

$$= \frac{2r_{O1}C_X s + 1}{2r_{O1}C_X s + 2g_{m2}r_{O1} + 1} g_{m2}R_1. \tag{5.72}$$

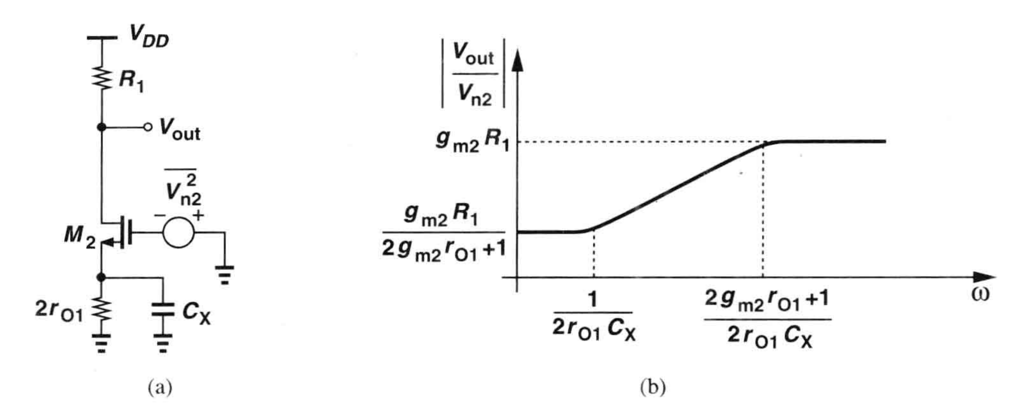


Figure 5.22 *(a) Cascode transistor noise, (b) output contribution as a function of frequency.*

Figure 5.22(b) plots the frequency response, implying that the noise contribution of M_2 is negligible for frequencies up to the zero frequency, $(2r_{O1}C_X)^{-1}$, but begins to manifest itself thereafter. Since C_X is comparable with C_{GS} and $2r_{O1} \gg 1/g_m$, we note that $(2r_{O1}C_X)^{-1} \ll g_m/C_{GS}$ $(\approx \omega_T)$. That is, the zero frequency is much lower than the f_T of the transistors, making this effect potentially significant.

Example 5.10

Assuming $2r_{O1} \gg |C_X s|^{-1}$ at frequencies of interest so that the degeneration impedance in the source of M_2 reduces to C_X, recompute the above transfer function while taking C_{GS2} into account. Neglect the effect of r_{O2}.

Solution:

From the equivalent circuit shown in Fig. 5.23, we have $g_{m2}V_1 = -V_{out}/R_1$ and hence $V_1 = -V_{out}/(g_{m2}R_1)$. The current flowing through C_{GS2} is therefore equal to $-V_{out}C_{GS2}s/(g_{m2}R_1)$. The sum of this current and $-V_{out}/R_1$ flows through C_X, producing a voltage of $[-V_{out}C_{GS2}s/(g_{m2}R_1) - V_{out}/R_1]/(C_X s)$. Writing a KVL in the input loop gives

$$\left(-\frac{V_{out}C_{GS2}s}{g_{m2}R_1} - \frac{V_{out}}{R_1}\right)\frac{1}{C_X s} - \frac{V_{out}}{g_{m2}R_1} = V_{in} \quad (5.73)$$

Figure 5.23 *Computation of gain from the gate of cascode device to output.*

Example 5.10 (*Continued*)

and hence

$$\frac{V_{out}}{V_{in}} = \frac{-g_{m2}R_1C_Xs}{(C_{GS2}+C_X)s+g_{m2}}. \tag{5.74}$$

At frequencies well below the f_T of the transistor, $|(C_{GS2}+C_X)s| \ll g_{m2}$ and

$$\frac{V_{out}}{V_{in}} \approx -R_1C_Xs. \tag{5.75}$$

That is, the noise of M_2 reaches the output unattenuated if ω is much greater than $(2r_{O1}C_X)^{-1}$ [but much less than $g_{m2}/(C_{GS2}+C_X)$].

The second issue stemming from the cascode device relates to the limited voltage headroom. To quantify this limitation, let us determine the required or allowable values of V_{b1} and V_{b2} in Fig. 5.21. Since the drain voltage of M_2 begins at V_{DD} and can swing below its gate voltage by as much as V_{TH2} while keeping M_2 in saturation, we can simply choose $V_{b2} = V_{DD}$. Now, $V_X = V_{DD} - V_{GS2}$, allowing a maximum value of $V_{DD} - V_{GS2} + V_{TH1}$ for V_{b1} if M_1 must remain saturated. Consequently, the source voltage of M_1 cannot exceed $V_{DD} - V_{GS2} - (V_{GS1} - V_{TH1})$. We say the two transistors consume a voltage headroom of one V_{GS} plus one overdrive $(V_{GS1} - V_{TH1})$.

It may appear that, so long as $V_{DD} > V_{GS2} + (V_{GS1} - V_{TH1})$, the circuit can be properly biased, but how about the path from the source of M_1 to ground? In comparison with the CG stages in Fig. 5.17, the cascode topology consumes an additional voltage headroom of $V_{GS1} - V_{TH1}$, leaving less for the biasing transistor or resistor and hence raising their noise contribution. For example, suppose $I_{D1} = I_{D2} = 2$ mA. Since $g_{m1} = (50\,\Omega)^{-1} = 2I_D/(V_{GS1} - V_{TH1})$, we have $V_{GS1} - V_{TH1} = 200$ mV. Also assume $V_{GS2} \approx 500$ mV. Thus, with $V_{DD} = 1$ V, the voltage available for a bias resistor, R_B, tied between the source of M_1 and ground cannot exceed $300\,\text{mV}/2\,\text{mA} = 150\,\Omega$. This value is comparable with $R_S = 50\,\Omega$ and degrades the gain and noise behavior of the circuit considerably.

In order to avoid the noise-headroom trade-off imposed by R_B, and also cancel the input capacitance of the circuit, CG stages often employ an inductor for the bias path. Illustrated in Fig. 5.24 with proper biasing for the input transistor, this technique minimizes the additional noise due to the biasing element (L_B) and significantly improves the input matching. In modern RF design, both L_B and L_1 are integrated on the chip.

Design Procedure With so many devices present in the circuit of Fig. 5.24, how do we begin the design? We describe a systematic procedure that provides a "first-order" design, which can then be refined and optimized.

The design procedure begins with two knowns: the frequency of operation and the supply voltage. In the first step, the dimensions and bias current of M_1 must be chosen such that a transconductance of $(50\,\Omega)^{-1}$ is obtained. The length of the transistor is set to the minimum allowable by the technology, but how should the width and the drain current be determined?

Using circuit simulations, we plot the transconductance and f_T of an NMOS transistor with a given width, W_0, as a function of the drain current. For long-channel devices,

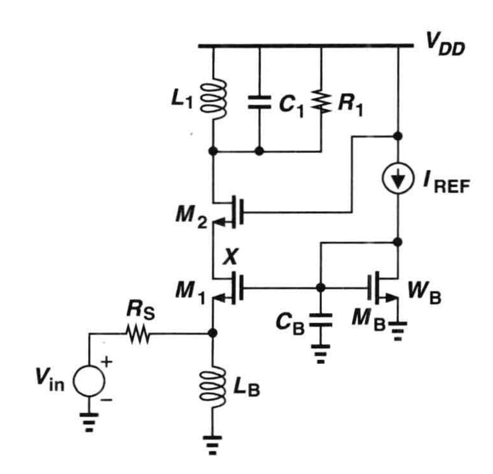


Figure 5.24 *Biasing of cascode CG stage.*

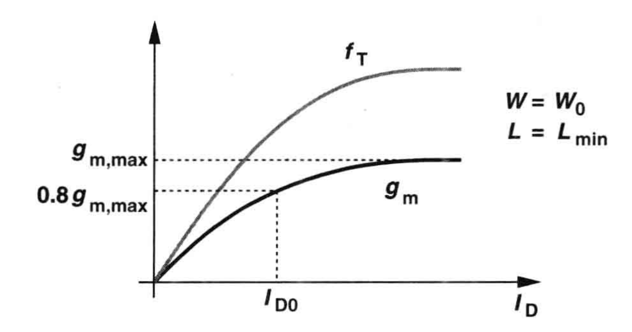


Figure 5.25 *Behavior of g_m and f_T as a function of drain current.*

$g_m \propto \sqrt{I_D}$, but submicron transistors suffer from degradation of the mobility with the vertical field in the channel, exhibiting the saturation behavior shown in Fig. 5.25. To avoid excessive power consumption, we select a bias current, I_{D0}, that provides 80 to 90% of the saturated g_m. That is, the combination of W_0 and I_{D0} (the "current density") is nearly optimum in terms of speed (transistor capacitances) and power consumption.

With W_0 and I_{D0} known, any other value of transconductance can be obtained by simply *scaling* the two proportionally. The reader can prove that if W_0 and I_{D0} scale by a factor of α, then so does g_m, regardless of the type and behavior of the transistor. We thus arrive at the required dimensions and bias current of M_1 (for $1/g_{m1} = 50\,\Omega$), which in turn yield its overdrive voltage.

In the second step, we compute the necessary value of L_B in Fig. 5.24. As shown in Fig. 5.26, the input of the circuit sees a pad capacitance to the substrate.[6] Thus, L_B must resonate with $C_{pad} + C_{SB1} + C_{GS1}$ and *its own capacitance* at the frequency of interest. (Here, R_p models the loss of L_B.) Since the parasitic capacitance of L_B is not known a priori, some iteration is required. (The design and modeling of spiral inductors are described in Chapter 7.)

6. The input may also see additional capacitance due to electrostatic discharge (ESD) protection devices that are tied to V_{DD} and ground.

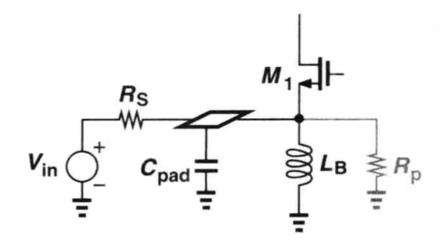


Figure 5.26 *Effect of pad capacitance on CG stage.*

Does L_B affect the performance of the circuit at resonance? Accompanying L_B is the parallel equivalent resistance $R_p = QL_B\omega$, which contributes noise and possibly attenuates the input signal. Thus, R_p must be at least ten times higher than $R_S = 50\,\Omega$. In other words, if the total capacitance at the input is so large as to dictate an excessively small inductor and R_p, then the noise figure is quite high. This situation may arise only at frequencies approaching the f_T of the technology.

In the third step, the bias of M_1 is defined by means of M_B and I_{REF} in Fig. 5.24. For example, $W_B = 0.2W_1$ and $I_{REF} = 0.2I_{D1}$ so that the bias branch draws only one-fifth of the current of the main branch.[7] Capacitor C_B provides a sufficiently low impedance (much less than $50\,\Omega$) from the gate of M_1 to ground and also bypasses the noise of M_B and I_B to ground. The choice of a solid, low-inductance ground is critical here because the high-frequency performance of the CG stage degrades drastically if the impedance seen in series with the gate becomes comparable with R_S.

Next, the width of M_2 in Fig. 5.24 must be chosen (the length is the minimum allowable value). With the bias current known ($I_{D2} = I_{D1}$), if the width is excessively small, then V_{GS2} may be so large as to drive M_1 into the triode region. On the other hand, as W_2 increases, M_2 contributes an increasingly larger capacitance to node X while its g_m reaches a nearly constant value (why?). Thus, the optimum width of M_2 is likely to be near that of M_1, and that is the initial choice. Simulations can be used to refine this choice, but in practice, even a twofold change from this value negligibly affects the performance.

In order to minimize the capacitance at node X in Fig. 5.24, transistors M_1 and M_2 can be laid out such that the drain area of the former is shared with the source area of the latter. Furthermore, since no other connection is made to this node, the shared area need not accommodate contacts and can therefore be minimized. Depicted in Fig. 5.27 and feasible only if $W_1 = W_2$, such a structure can be expanded to one with multiple gate fingers.

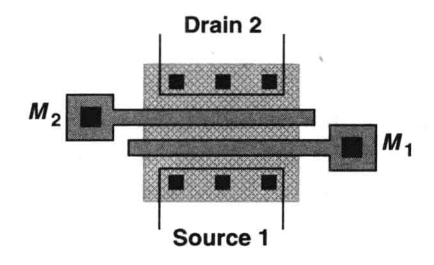


Figure 5.27 *Layout of cascode devices.*

7. For proper matching between the two transistors, M_1 incorporates five unit transistors (e.g., gate fingers) and M_B one unit transistor.

In the last step, the value of the load inductor, L_1, must be determined (Fig. 5.24). In a manner similar to the choice of L_B, we compute L_1 such that it resonates with $C_{GD2} + C_{DB2}$, the input capacitance of the next stage, and its own capacitance. Since the voltage gain of the LNA is proportional to $R_1 = QL_1\omega$, R_1 must be sufficiently large, e.g., 500 to 1000 Ω. This condition is met in most designs without much difficulty.

The design procedure outlined above leads to a noise figure around 3 dB [Eq. (5.58)] and a voltage gain, $V_{out}/V_{in} = R_1/(2R_S)$, of typically 15 to 20 dB. If the gain is *too high*, i.e., if it dictates an unreasonably high mixer IP_3, then an explicit resistor can be placed in parallel with R_1 to obtain the required gain. As studied in Section 5.7, this LNA topology displays a high IIP_3, e.g., +5 to +10 dBm.

Example 5.11

Design the LNA of Fig. 5.24 for a center frequency of 5.5 GHz in 65-nm CMOS technology. Assume the circuit is designed for an 11a receiver.

Solution:

Figure 5.28 plots the transconductance of an NMOS transistor with $W = 10\ \mu$m and $L = 60$ nm as a function of the drain current. We select a bias current of 2 mA to achieve a g_m of about 10 mS $= 1/(100\ \Omega)$. Thus, to obtain an input resistance of 50 Ω, we must double the width and drain current.[8] The capacitance introduced by a 20-μm transistor at the input is about 30 fF. To this we add a pad capacitance of 50 fF and choose $L_B = 10$ nH for resonance at 5.5 GHz. Such an inductor exhibits a parasitic capacitance of roughly 30 fF, requiring that a smaller inductance be chosen, but we proceed without this refinement.

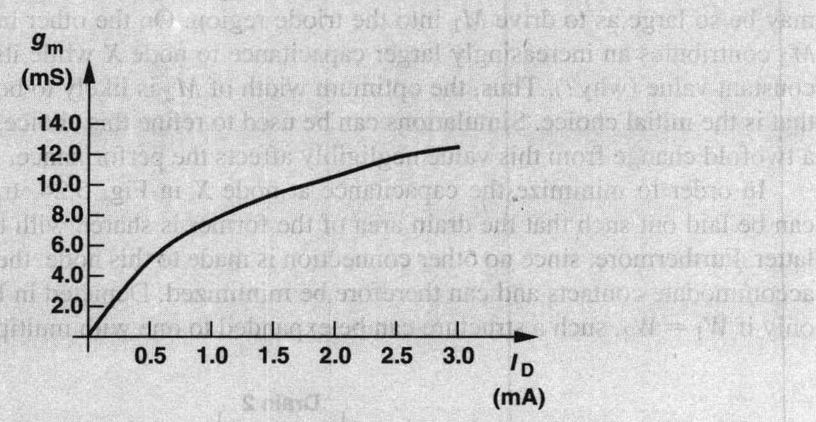


Figure 5.28 *Transcoductance of a 10 μm/60 nm NMOS device as a function of drain current.*

Next, we choose the width of the cascode device equal to 20 μm and assume a load capacitance of 30 fF (e.g., the input capacitance of subsequent mixers). This allows the use of a 10-nH inductor for the load, too, because the total capacitance at the output node amounts to about 75 fF. However, with a Q of about 10 for such an inductor, the LNA

8. The body effect lowers the input resistance, but the feedback from the drain to the gate raises it. We therefore neglect both.

Example 5.11 (*Continued*)

gain is exessively high and its bandwidth excessively low (failing to cover the 11a band). For this reason, we place a resistor of 1 kΩ in parallel with the tank. Figure 5.29 shows the design details and Fig. 5.30 the simulated characteristics. Note that the inductor loss

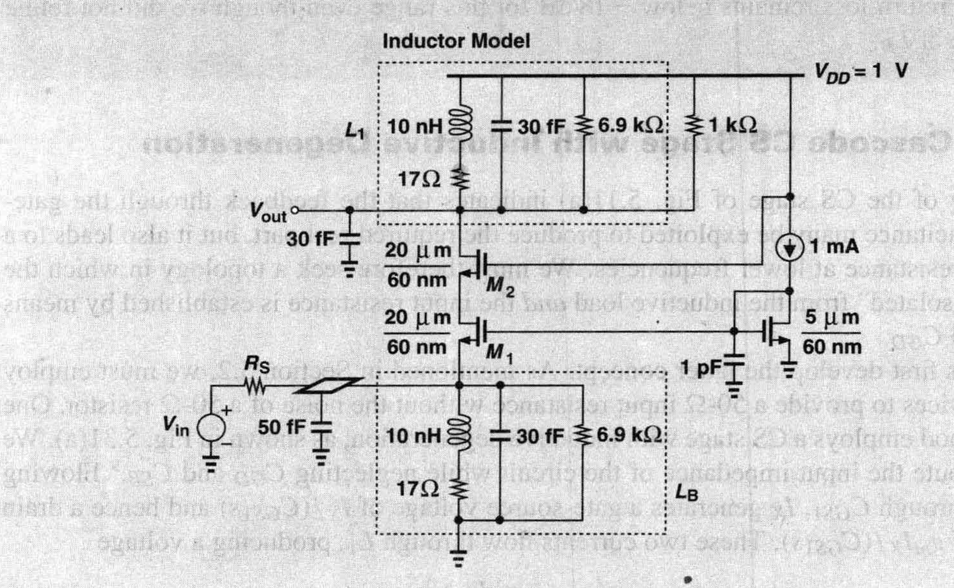


Figure 5.29 *CG LNA example.*

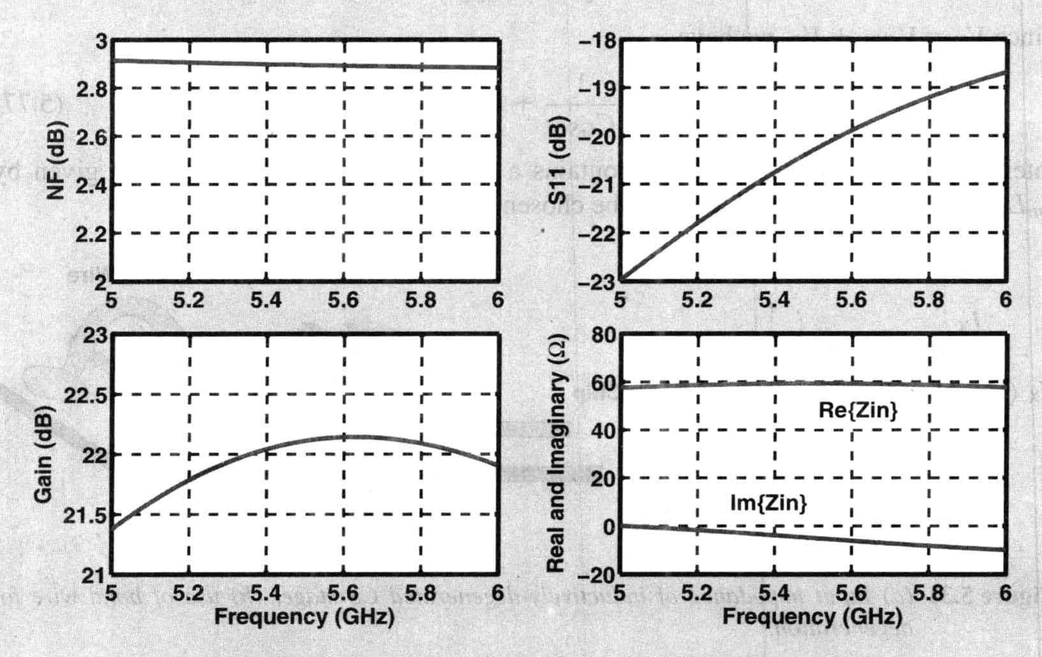


Figure 5.30 *Simulated charactersitics of CG LNA example.*

(*Continues*)

Example 5.11 *(Continued)*

is modeled by series and parallel resistances so as to obtain a broadband representation (Chapter 7).

The simulation results reveal a relatively flat noise figure and gain from 5 to 6 GHz. The input return loss remains below −18 dB for this range even though we did not refine the choice of L_B.

5.3.4 Cascode CS Stage with Inductive Degeneration

Our study of the CS stage of Fig. 5.11(a) indicates that the feedback through the gate-drain capacitance many be exploited to produce the required real part, but it also leads to a negative resistance at lower frequencies. We must therefore seek a topology in which the input is "isolated" from the inductive load *and* the input resistance is established by means other than C_{GD}.

Let us first develop the latter concept. As mentioned in Section 5.2, we must employ active devices to provide a 50-Ω input resistance without the noise of a 50-Ω resistor. One such method employs a CS stage with inductive degeneration, as shown in Fig. 5.31(a). We first compute the input impedance of the circuit while neglecting C_{GD} and C_{SB}.[9] Flowing entirely through C_{GS1}, I_X generates a gate-source voltage of $I_X/(C_{GS1}s)$ and hence a drain current of $g_m I_X/(C_{GS1}s)$. These two currents flow through L_1, producing a voltage

$$V_P = \left(I_X + \frac{g_m I_X}{C_{GS1}s}\right) L_1 s. \tag{5.76}$$

Since $V_X = V_{GS1} + V_P$, we have

$$\frac{V_X}{I_X} = \frac{1}{C_{GS1}s} + L_1 s + \frac{g_m L_1}{C_{GS1}}. \tag{5.77}$$

Interestingly, the input impedance contains a frequency-independent real part given by $g_m L_1/C_{GS1}$. Thus, the third term can be chosen equal to 50 Ω.

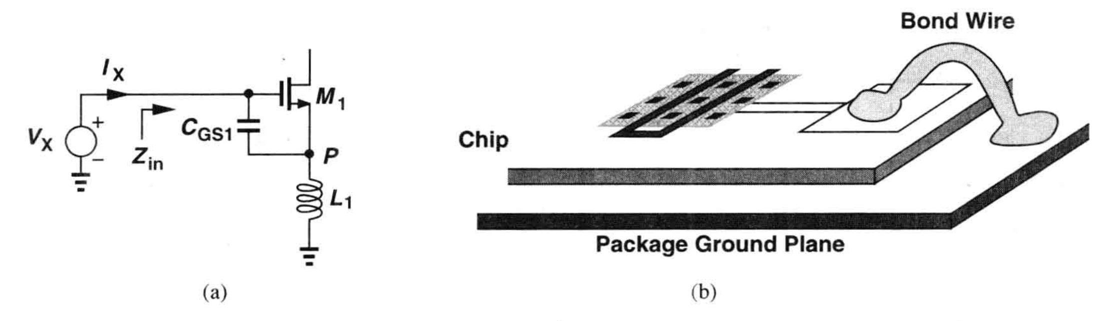


Figure 5.31 *(a) Input impedance of inductively-degenerated CS stage, (b) use of bond wire for degeneration.*

9. We also neglect channel-length modulation and body effect.

The third term in Eq. (5.77) carries a profound meaning: since $g_m/C_{GS1} \approx \omega_T\ (=2\pi f_T)$, the input resistance is approximately equal to $L_1\omega_T$ and directly related to the f_T of the transistor. For example, in 65-nm technology, $\omega_T \approx 2\pi \times (160\,\text{GHz})$, dictating $L_1 \approx 50\,\text{pH}$ (!) for a real part of 50 Ω.

In practice, the degeneration inductor is often realized as a bond wire with the reasoning that the latter is inevitable in packaging and must be incorporated in the design. To minimize the inductance, a "downbond" can directly connect the source pad to a ground plane in the package [Fig. 5.31(b)], but even this geometry yields a value in the range of 0.5 to 1 nH—far from the 50-pH amount calculated above! That is, the input resistance provided by modern MOSFETs tends to be substantially higher than 50 Ω if a bond wire inductance is used.[10]

How do we obtain a 50-Ω resistance with $L_1 \approx 0.5$ nH? At operation frequencies far below f_T of the transistor, we can *reduce* the f_T. This is accomplished by increasing the channel length or simply placing an explicit capacitor in parallel with C_{GS}. For example, if $L_1 = 0.5$ nH, then f_T must be lowered to about 16 GHz.

Example 5.12

Determine the input impedance of the circuit shown in Fig. 5.32(a) if C_{GD} is not neglected and the drain is tied to a load resistance R_1. Assume $R_1 \approx 1/g_m$ (as in a cascode).

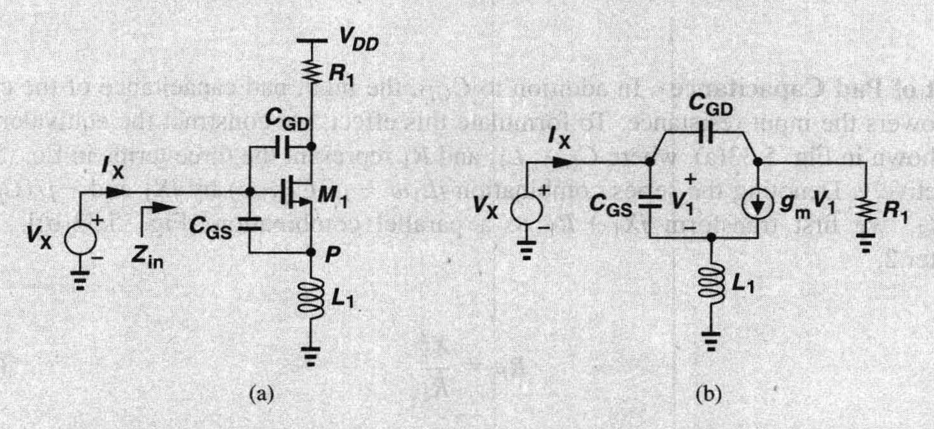


Figure 5.32 *(a) Input impedance of CS stage in the presence of C_{GD}, (b) equivalent circuit.*

Solution:

From the equivalent circuit depicted in Fig. 5.32(b), we note the current flowing through L_1 is equal to $V_1C_{GS}s + g_mV_1$ and hence

$$V_X = V_1 + (V_1C_{GS}s + g_mV_1)L_1s. \tag{5.78}$$

(Continues)

10. This is a rare case in which the transistor is too fast!

Example 5.12 *(Continued)*

Also, the current flowing through R_1 is equal to $I_X - V_1 C_{GS}s - g_m V_1$, leading to

$$V_X = (I_X - V_1 C_{GS}s - g_m V_1)R_1 + (I_X - V_1 C_{GS}s)\frac{1}{C_{GD}s}. \tag{5.79}$$

Substituting for V_1 from (5.78), we have

$$\frac{V_X}{I_X} = \frac{\left(R_1 + \frac{1}{C_{GD}s}\right)(L_1 C_{GS}s^2 + g_m L_1 s + 1)}{L_1 C_{GS}s^2 + (R_1 C_{GS} + g_m L_1)s + g_m R_1 + C_{GS}/C_{GD} + 1}. \tag{5.80}$$

If $R_1 \approx 1/g_m \ll |C_{GD}s|^{-1}$ and C_{GS}/C_{GD} dominates in the denominator, (5.80) reduces to

$$\frac{V_X}{I_X} \approx \left(\frac{1}{C_{GS}s} + L_1 s + \frac{g_m L_1}{C_{GS}}\right)\left[1 - \frac{2C_{GD}}{C_{GS}} - L_1 C_{GD}s^2 - \left(R_1 C_{GD} + g_m L_1 \frac{C_{GD}}{C_{GS}}\right)s\right] \tag{5.81}$$

Assuming that the first two terms in the square brackets are dominant, we conclude that the input resistance falls by a factor of $1 - 2C_{GD}/C_{GS}$.

Effect of Pad Capacitance In addition to C_{GD}, the input pad capacitance of the circuit also lowers the input resistance. To formulate this effect, we construct the equivalent circuit shown in Fig. 5.33(a), where C_{GS1}, L_1, and R_1 represent the three terms in Eq. (5.77), respectively. Denoting the series combination $jL_1\omega - j/(C_{GS1}\omega)$ by jX_1 and $-j/(C_{pad}\omega)$ by jX_2, we first transform $jX_1 + R_1$ to a parallel combination [Fig. 5.33(b)]. From Chapter 2,

$$R_P = \frac{X_1^2}{R_1}. \tag{5.82}$$

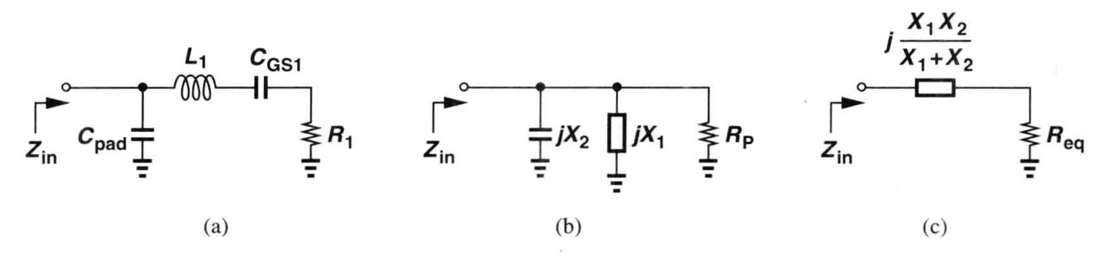


Figure 5.33 *(a) Equivalent circuit for inclusion of pad capacitance, (b) simplified circuit of (a), (c) simplified circuit of (b).*

We now merge the two parallel reactances into $jX_1X_2/(X_1+X_2)$ and transform the resulting circuit to a series combination [Fig. 5.33(c)], where

$$R_{eq} = \left(\frac{X_1X_2}{X_1+X_2}\right)^2 \cdot \frac{1}{R_P} \tag{5.83}$$

$$= \left(\frac{X_2}{X_1+X_2}\right)^2 R_1. \tag{5.84}$$

In most cases, we can assume $L_1\omega \ll 1/(C_{GS1}\omega) + 1/(C_{pad}\omega)$ at the frequency of interest, obtaining

$$R_{eq} \approx \left(\frac{C_{GS1}}{C_{GS1}+C_{pad}}\right)^2 R_1. \tag{5.85}$$

For example, if $C_{GS1} \approx C_{pad}$, then the input resistance falls by a factor of four.

We can now make two observations. First, the effect of the gate-drain and pad capacitance suggests that the transistor f_T need not be reduced so much as to create $R_1 = 50\ \Omega$. Second, since the degeneration inductance necessary for $Re\{Z_{in}\} = 50\ \Omega$ is insufficient to resonate with $C_{GS1} + C_{pad}$, another inductor must be placed in series with the gate as shown in Fig. 5.34, where it is assumed L_G is off-chip.

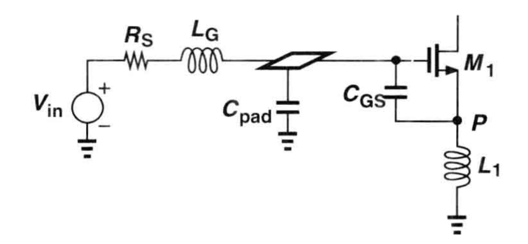


Figure 5.34 *Addition of L_G for input matching.*

Example 5.13

A 5-GHz LNA requires a value of 2 nH for L_G. Discuss what happens if L_G is integrated on the chip and its Q does not exceed 5.

Solution:

With $Q = 5$, L_G suffers from a series resistance equal to $L_G\omega/Q = 12.6\ \Omega$. This value is not much less than $50\ \Omega$, degrading the noise figure considerably. For this reason, L_G is typically placed off-chip.

NF Calculation Let us now compute the noise figure of the CS circuit, excluding the effect of channel-length modulation, body effect, C_{GD}, and C_{pad} for simplicity (Fig. 5.35). The noise of M_1 is represented by I_{n1}. For now, we assume the output of interest is the

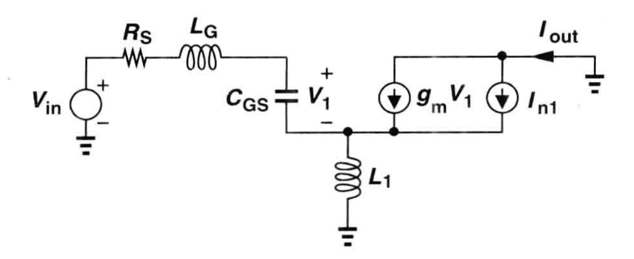


Figure 5.35 *Equivalent circuit for computation of NF.*

current I_{out}. We have

$$I_{out} = g_m V_1 + I_{n1}. \tag{5.86}$$

Also, since L_1 sustains a voltage of $L_1 s(I_{out} + V_1 C_{GS1} s)$, a KVL around the input loop yields

$$V_{in} = (R_S + L_G s)V_1 C_{GS1} s + V_1 + L_1 s(I_{out} + V_1 C_{GS1} s). \tag{5.87}$$

Substituting for V_1 from (5.86) gives

$$V_{in} = I_{out} L_1 s + \frac{(L_1 + L_G)C_{GS1}s^2 + 1 + R_S C_{GS1} s}{g_m}(I_{out} - I_{n1}). \tag{5.88}$$

The input network is designed to resonate at the frequency of interest, ω_0. That is, $(L_1 + L_G)C_{GS1} = \omega_0^{-2}$ and hence, $(L_1 + L_G)C_{GS1}s^2 + 1 = 0$ at $s = j\omega_0$. We therefore obtain

$$V_{in} = I_{out}\left(jL_1\omega_0 + \frac{jR_S C_{GS1}\omega_0}{g_m}\right) - I_{n1}\frac{jR_S C_{GS1}\omega_0}{g_m}. \tag{5.89}$$

The coefficient of I_{out} represents the transconductance gain of the circuit (including R_S):

$$\left|\frac{I_{out}}{V_{in}}\right| = \frac{1}{\omega_0\left(L_1 + \frac{R_S C_{GS1}}{g_m}\right)}. \tag{5.90}$$

Now, recall from Eq. (5.77) that, for input matching, $g_m L_1 / C_{GS1} = R_S$. Since $g_m / C_{GS1} \approx \omega_T$,

$$\left|\frac{I_{out}}{V_{in}}\right| = \frac{\omega_T}{2\omega_0} \cdot \frac{1}{R_S}. \tag{5.91}$$

Interestingly, the transconductance of the circuit remains independent of L_1, L_G, and g_m so long as the input is matched.

Setting V_{in} to zero in Eq. (5.89), we compute the output noise due to M_1:

$$|I_{n,out}|_{M1} = |I_{n1}|\frac{R_S C_{GS1}}{g_m L_1 + R_S C_{GS1}}, \tag{5.92}$$

which, for $g_m L_1/C_{GS1} = R_S$, reduces to

$$|I_{n,out}|_{M1} = \frac{|I_{n1}|}{2}, \tag{5.93}$$

and hence

$$\overline{I_{n,out}^2}|_{M1} = kT\gamma g_m. \tag{5.94}$$

Dividing the output noise current by the transconductance of the circuit and by $4kTR_S$ and adding unity to the result, we arrive at the noise figure of the circuit [2]:

$$\text{NF} = 1 + g_m R_S \gamma \left(\frac{\omega_0}{\omega_T}\right)^2. \tag{5.95}$$

It is important to bear in mind that this result holds only at the input resonance frequency and if the input is matched.

Example 5.14

A student notes from Eq. (5.95) above that, if the transistor width and bias current are scaled *down* proportionally, then g_m and C_{GS1} decrease while $g_m/C_{GS1} = \omega_T$ remains constant. That is, the noise figure decreases while the power dissipation of the circuit also decreases! Does this mean we can obtain NF = 1 with zero power dissipation?

Solution:

As C_{GS1} decreases, $L_G + L_1$ must increase proportionally to maintain a constant ω_0. Suppose L_1 is fixed and we simply increase L_G. As C_{GS1} approaches zero and L_G infinity, the Q of the input network ($\approx L_G\omega_0/R_S$) also goes to infinity, providing an *infinite voltage gain at the input*. Thus, the noise of R_S overwhelms that of M_1, leading to NF = 1. This result is not surprising; after all, in a circuit such as the network of Fig. 5.36, $|V_{out}/V_{in}| = (R_S C_a \omega_0)^{-1}$ at resonance, implying that the voltage gain approaches infinity if C_a goes to zero (and L_a goes to infinity so that ω_0 is constant). In practice, of course, the inductor suffers from a finite Q (and parasitic capacitances), limiting the performance.

Figure 5.36 *Equivalent circuit of CS input network.*

What if we keep L_G constant and increase the degeneration inductance, L_1? The NF still approaches 1 but the transconductance of the circuit, Eq. (5.90), falls to zero if C_{GS1}/g_m remains fixed.[11] That is, the circuit provides a zero-dB noise figure but with zero gain.

11. If C_{GS1}/g_m is constant and L_1 increases, the input cannot remain matched and Eq. (5.95) is invalid.

The above example suggests that maximizing L_G can minimize the noise figure by providing voltage *gain* from V_{in} to the gate of M_1. The reader can prove that this gain is given by

$$\frac{V_G}{V_{in}} = \frac{1}{2}\left(1 + \frac{L_G\omega_0}{R_S}\right). \tag{5.96}$$

Note that $L_G\omega_0/R_S$ represents the Q of the series combination of L_G and R_S. Indeed, as explained below, the design procedure begins with the maximum available value of L_G (typically an off-chip inductor) whose parasitic capacitances are negligible. The voltage gain in the input network (typically as high as 6 dB) does lower the IP_3 and P_{1dB} of the LNA, but the resulting values still prove adequate in most applications.

We now turn our attention to the output node of the circuit. As explained in Section 5.3.1, an inductive load attached to a common-source stage introduces a negative resistance due to the feedback through C_{GD}. We therefore add a cascode transistor in the output branch to suppress this effect. Figure 5.37 shows the resulting circuit, where R_1 models the loss of L_D. The voltage gain is equal to the product of the circuit's transconductance [Eq. (5.91)] and the load resistance, R_1:[12]

$$\frac{V_{out}}{V_{in}} = \frac{\omega_T}{2\omega_0}\frac{R_1}{R_S} \tag{5.97}$$

$$= \frac{R_1}{2L_1\omega_0}. \tag{5.98}$$

The effect of C_{GD1} on the input impedance may still require attention because the impedance seen at the source of M_2, R_X, rises sharply at the output resonance frequency. From Eq. (5.64),

$$R_X = \frac{R_1 + r_{O2}}{1 + g_m r_{O2}}. \tag{5.99}$$

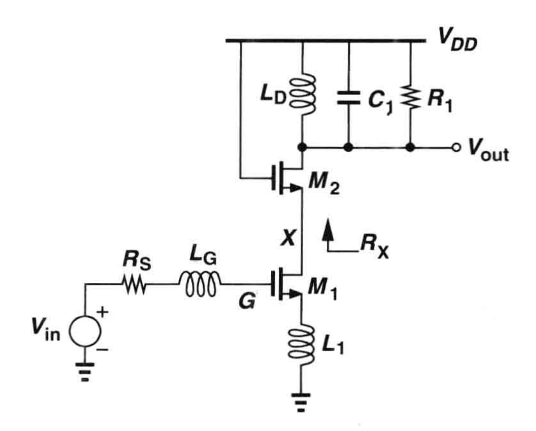


Figure 5.37 *Inductively-degenerated cascode CS LNA.*

12. The output impedance of the cascode is assumed much higher than R_1.

Using the transconductance expression in (5.90) and V_G/V_{in} in (5.96), we compute the voltage gain from the gate to the drain of M_1:

$$\frac{V_X}{V_G} = \frac{R_S}{L_1\omega_0} \cdot \frac{R_1 + r_{O2}}{(1 + g_{m2}r_{O2})(R_S + L_G\omega_0)}. \tag{5.100}$$

Since $R_S \gg L_1\omega_0$ (why?) and the second fraction is typically near or higher than unity, C_{GD1} may suffer from substantial Miller multiplication at the output resonance frequency.

In the foregoing noise figure calculation, we have not included the noise contribution of M_2. As formulated for the cascode CG stage in Section 5.3.3, the noise of the cascode device begins to manifest itself if the frequency of operation exceeds roughly $(2r_{O1}C_X)^{-1}$.

Example 5.15

Determine the noise figure of the cascode CS stage of Fig. 5.37, including the noise contributed by R_1 but neglecting the noise of M_2.

Solution:

Dividing the noise of R_1 by the gain given by (5.98) and the noise of R_S and adding the result to the noise figure in (5.95), we have

$$NF = 1 + g_m R_S \gamma \left(\frac{\omega_0}{\omega_T}\right)^2 + \frac{4R_S}{R_1}\left(\frac{\omega_0}{\omega_T}\right)^2. \tag{5.101}$$

Design Procedure Having developed a good understanding of the cascode CS LNA of Fig. 5.37, we now describe a procedure for designing the circuit. The reader is encouraged to review the CG design procedure. The procedure begins with four knowns: the frequency of operation, ω_0, the value of the degeneration inductance, L_1, the input pad capacitance, C_{pad}, and the value of the input series inductance, L_G. Each of the last three knowns is somewhat flexible, but it is helpful to select some values, complete the design, and iterate if necessary.

Governing the design are the following equations:

$$\frac{1}{(L_G + L_1)(C_{GS1} + C_{pad})} = \omega_0^2 \tag{5.102}$$

$$\left(\frac{C_{GS1}}{C_{GS1} + C_{pad}}\right)^2 L_1\omega_T = R_S. \tag{5.103}$$

With ω_0 known, C_{GS1} is calculated from (5.102), and ω_T and g_m ($= \omega_T C_{GS1}$) from (5.103). We then return to the plots of g_m and f_T in Fig. 5.25 and determine whether a transistor width can yield the necessary g_m and f_T simultaneously. In deep-submicron technologies and for operation frequencies up to a few tens of gigahertz, the f_T is likely to be "too high," but the pad capacitance alleviates the issue by transforming the input resistance to a lower value. If the requisite f_T is quite low, a capacitance can be added to C_{pad}. On the other hand, if the pad capacitance is so large as to demand a very high f_T, the degeneration inductance can be increased.

In the next step, the dimensions of the cascode device are chosen equal to those of the input transistor. As mentioned in Section 5.3.3 for the cascode CG stage, the width of the cascode device only weakly affects the performance. Also, the layout of M_1 and M_2 can follow the structure shown in Fig. 5.27 to minimize the capacitance at node X.

The design procedure now continues with selecting a value for L_D such that it resonates at ω_0 with the drain-bulk and drain-gate capacitances of M_2, the input capacitance of the next stage, and the inductors's own parasitic capacitance. If the parallel equivalent resistance of L_D results in a gain, $R_1/(2L_1\omega_0)$, greater than required, then an explicit resistor can be placed in parallel with L_D to lower the gain and widen the bandwidth.

In the last step of the design, we must examine the input match. Due to the Miller multiplication of C_{GD1} (Example 5.12), it is possible that the real and imaginary parts depart from their ideal values, necessitating some adjustment in L_G.

The foregoing procedure typically leads to a design with a relatively low noise figure, around 1.5 to 2 dB—depending on how large L_G can be without displaying excessive parasitic capacitances. Alternatively, the design procedure can begin with known values for NF and L_1 and the following two equations:

$$\mathrm{NF} = 1 + g_{m1}R_S\gamma\left(\frac{\omega_0}{\omega_T}\right)^2 \tag{5.104}$$

$$R_S = \left(\frac{C_{GS1}}{C_{GS1} + C_{pad}}\right)^2 L_1\omega_T, \tag{5.105}$$

where the noise of the cascode transistor and the load is neglected. The necessary values of ω_T and g_{m1} can thus be computed ($g_{m1}/C_{GS1} \approx \omega_T$). If the plots in Fig. 5.25 indicate that the device f_T is too high, then additional capacitance can be placed in parallel with C_{GS1}. Finally, L_G is obtained from Eq. (5.102). (If advanced packaging minimizes inductances, then L_1 can be integrated on the chip and assume a small value.)

The overall LNA appears as shown in Fig. 5.38, where the antenna is capacitively tied to the receiver to isolate the LNA bias from external connections. The bias current of M_1 is

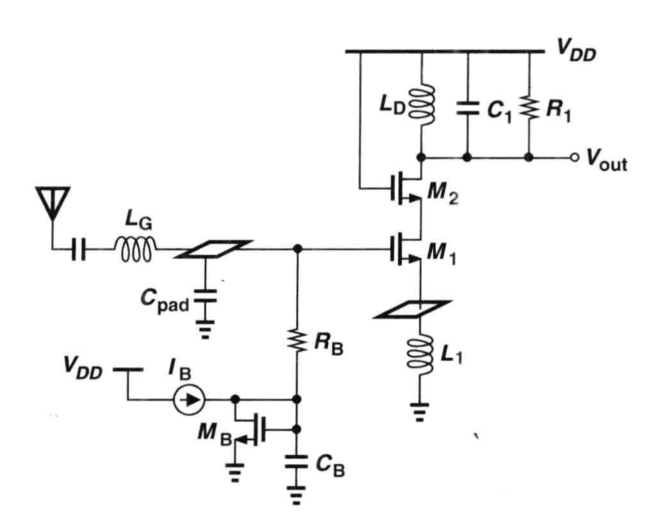


Figure 5.38 *Inductively-degenerated CS stage with pads and bias network.*

established by M_B and I_B, and resistor R_B and capacitor C_B isolate the signal path from the noise of I_B and M_B. The source-bulk capacitance of M_1 and the capacitance of the pad at the source of M_1 may slightly alter the input impedance and must be included in simulations.

Example 5.16

How is the value of R_B chosen in Fig. 5.38?

Solution:

Since R_B appears *in parallel* with the signal path, its value must be maximized. Is $R_B = 10R_S$ sufficiently high? As illustrated in Fig. 5.39, the series combination of R_S and L_G can be transformed to a parallel combination with $R_P \approx Q^2R_S \approx (L_G\omega_0/R_S)^2R_S$. From Eq. (5.96), we note that a voltage gain of, say, 2 at the input requires $Q = 3$, yielding $R_P \approx 450\,\Omega$. Thus, $R_B = 10R_S$ becomes *comparable* with R_P, raising the noise figure and lowering the voltage gain. In other words, R_B must remain much greater than R_P.

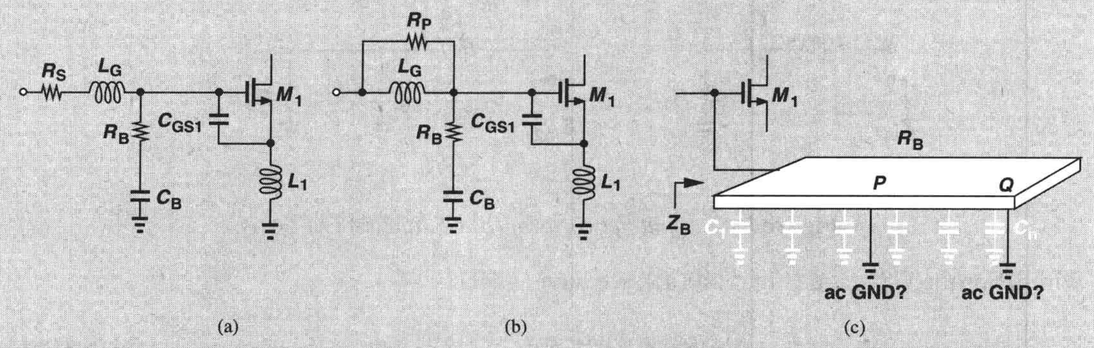


Figure 5.39 *(a) Effect of bias resistor R_B on CS LNA, (b) conversion of R_S and L_G to a parallel network, (c) effect of distributed capacitance of R_B.*

Large resistors may suffer from significant parasitic capacitance. However, increasing the *length* of a resistor does not load the signal path anymore even though it leads to a larger overall parasitic capacitance. To understand this point, consider the arrangement shown in Fig. 5.39(b), where the parasitic capacitance of R_B is represented as distributed elements C_1-C_n. Which node should be bypassed to ground, P or Q? We recognize that Z_B is *higher* if Q is bypassed even though the longer resistor has a higher capacitance. Thus, longer bias resistors are better. Alternatively, a small MOSFET acting as a resistor can be used here.

The choice between the CG and CS LNA topologies is determined by the trade-off between the robustness of the input match and the lower bound on the noise figure. The former provides an accurate input resistance that is relatively independent of package parasitics, whereas the latter exhibits a lower noise figure. We therefore select the CG stage if the required LNA noise figure can be around 4 dB, and the CS stage for lower values.

An interesting point of contrast between the CG and CS LNAs relates to the contribution of the load resistor, R_1, to the noise figure. Equation (5.58) indicates that in a CG stage, this contribution, $4R_S/R_1$, is equal to 4 divided by the voltage gain from the input

source to the output. Thus, for a typical gain of 10, this contribution reaches 0.4, a significant amount. For the inductively-degenerated CS stage, on the other hand, Eq. (5.101) reveals that the contribution is equal to $4R_S/R_1$ *multiplied* by $(\omega_0/\omega_T)^2$. Thus, for operation frequencies well below the f_T of the transistor, the noise contribution of R_1 becomes negligible.

Example 5.17

It is believed that input matching holds across a wider bandwidth for the CG stage than for the inductively-degenerated CS stage. Is this statement correct?

Solution:

Consider the equivalent circuits shown in Fig. 5.40 for the two LNA configurations, where $R_1 = 50\,\Omega$, C_1 and C_2 are roughly equal, and the inductors represent (inevitable) bond

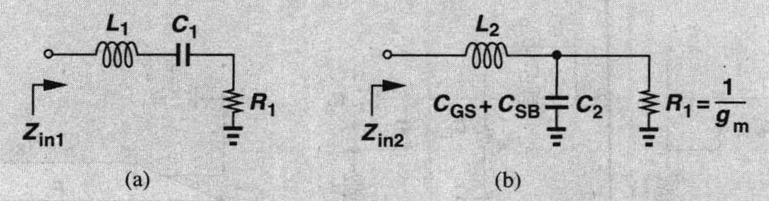


Figure 5.40 *Input networks of (a) CS and (b) CG LNAs.*

wires. For the CS stage [Fig. 5.40(a)], we have

$$Re\{Z_{in1}\} = R_1 \tag{5.106}$$

$$Im\{Z_{in1}\} = \frac{L_1C_1\omega^2 - 1}{C_1\omega}. \tag{5.107}$$

If the center frequency of interest is $\omega_0\ (= 1/\sqrt{L_1C_1})$ and $\omega = \omega_0 + \Delta\omega$, then

$$Im\{Z_{in1}\} \approx 2L_1\Delta\omega\frac{L_1\Delta\omega}{\omega_0}. \tag{5.108}$$

That is, the imaginary part varies in proportion to deviation from the center frequency, limiting the bandwidth across which $|S_{11}|$ remains acceptably low.

In the network of Fig. 5.40(b), on the other hand,

$$Re\{Z_{in2}\} = \frac{R_1}{1 + R_1^2C_2^2\omega^2} \tag{5.109}$$

$$Im\{Z_{in2}\} = L_2\omega - \frac{R_1^2C_2\omega}{1 + R_1^2C_2^2\omega^2}. \tag{5.110}$$

Example 5.17 (*Continued*)

In practice, $1/(R_1C_2)$ is comparable with the ω_T of the transistor [e.g., if $R_1 = 1/g_m$ and $C_2 = C_{GS}$, then $1/(R_1C_2) \approx \omega_T$]. Thus, for $\omega \ll \omega_T$,

$$Re\{Z_{in2}\} \approx R_1 \tag{5.111}$$

$$Im\{Z_{in2}\} \approx (L_2 - R_1^2C_2)\omega. \tag{5.112}$$

Interestingly, if $L_2 = R_1^2C_2$, then $Im\{Z_{in2}\}$ falls to zero and becomes *independent* of frequency. Thus the CG stage indeed provides a much broader band at its input, another advantage of this topology.

Example 5.18

Design a cascode CS LNA for a center frequency of 5.5 GHz in 65-nm CMOS technology.

Solution:

We begin with a degeneration inductance of 1 nH and the same input transistor as that in the CG stage of Example 5.11. Interestingly, with a pad capacitance of 50 fF, the input resistance happens to be around 60 Ω. (Without the pad capacitance, $Re\{Z_{in}\}$ is in the vicinity of 600 Ω.) We thus simply add enough inductance in series with the gate (L_G = 12 nH) to null the reactive component at 5.5 GHz. The design of the cascode device and the output network is identical to that of the CG example.

Figure 5.41 shows the details of the design and Fig. 5.42 the simulated characteristics. We observe that the CS stage has a higher gain, a lower noise figure, and a narrower bandwidth than the CG stage in Example 5.11.

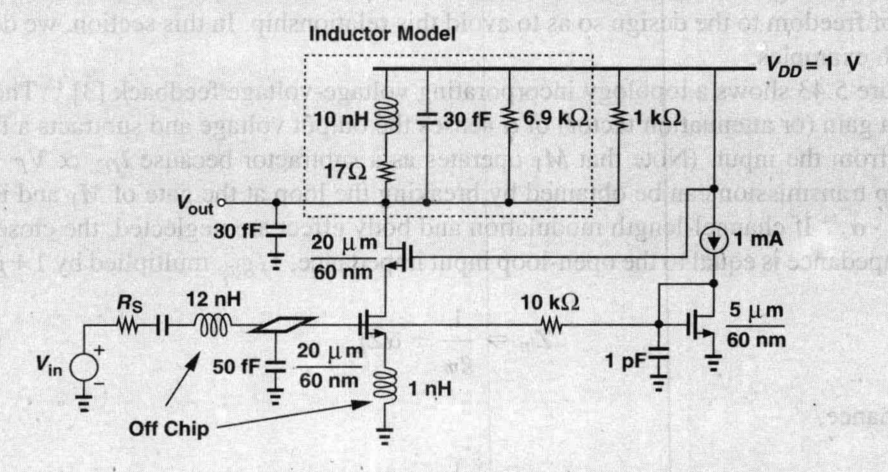


Figure 5.41 *CS LNA example.*

(*Continues*)

Example 5.18 *(Continued)*

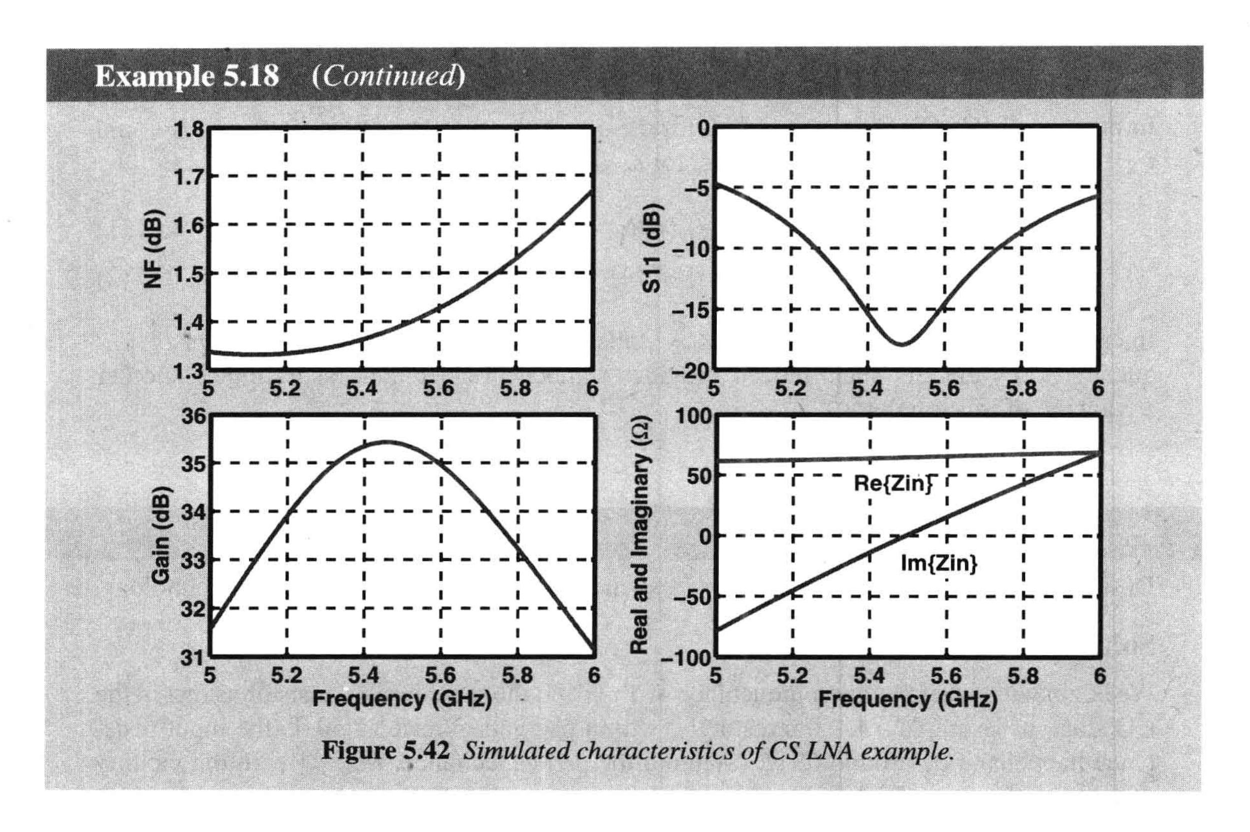


Figure 5.42 *Simulated characteristics of CS LNA example.*

5.3.5 Variants of Common-Gate LNA

As revealed by Eq. (5.57), the noise figure and input matching of the CG stage are inextricably related if channel-length modulation is negligible, a common situation in older CMOS technologies. For this reason, a number of efforts have been made to add another degree of freedom to the design so as to avoid this relationship. In this section, we describe two such examples.

Figure 5.43 shows a topology incorporating voltage-voltage feedback [3].[13] The block having a gain (or attenuation factor) of α senses the output voltage and subtracts a fraction thereof from the input. (Note that M_1 operates as a subtractor because $I_{D1} \propto V_F - V_{in}$.) The loop transmission can be obtained by breaking the loop at the gate of M_1 and is equal to $g_m Z_L \cdot \alpha$.[14] If channel-length modulation and body effect are neglected, the closed-loop input impedance is equal to the open-loop input impedance, $1/g_m$, multiplied by $1 + g_m Z_L \alpha$:

$$Z_{in} = \frac{1}{g_m} + \alpha Z_L. \tag{5.113}$$

At resonance,

$$Z_{in} = \frac{1}{g_m} + \alpha R_1. \tag{5.114}$$

13. This technique was originally devised for bipolar stages.
14. The input impedance of the feedback circuit is absorbed in Z_L.

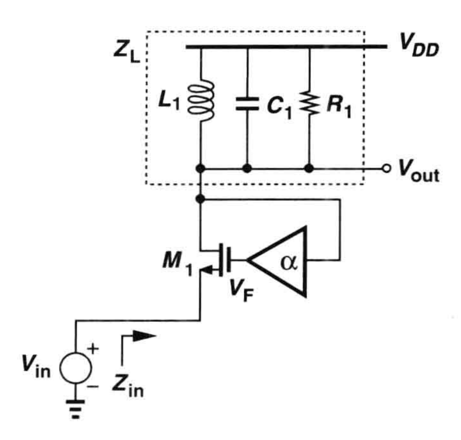


Figure 5.43 *CG LNA with feedback.*

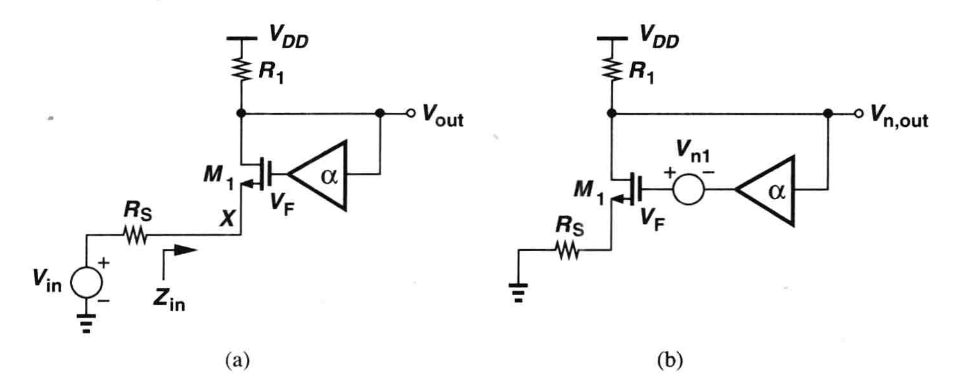


Figure 5.44 *(a) Input impedance and (b) noise behavior of CG stage with feedback.*

The input resistance can therefore be substantially *higher* than $1/g_m$, but how about the noise figure? We first calculate the gain with the aid of the circuit depicted in Fig. 5.44(a). The voltage gain from X to the output is equal to the open-loop gain, g_mR_1, divided by $1 + \alpha g_mR_1$ (at the resonance frequency). Thus,

$$\frac{V_{out}}{V_{in}} = \frac{Z_{in}}{Z_{in} + R_S} \cdot \frac{g_mR_1}{1 + \alpha g_mR_1} \tag{5.115}$$

$$= \frac{R_1}{\dfrac{1}{g_m} + \alpha R_1 + R_S}, \tag{5.116}$$

which reduces to $R_1/(2R_S)$ if the input is matched.

For output noise calculation, we construct the circuit of Fig. 5.44(b), where V_{n1} represents the noise voltage of M_1 and noise of the feedback circuit is neglected. Since R_S carries a current equal to $-V_{n,out}/R_1$ (why?), we recognize that $V_{GS1} = \alpha V_{n,out} + V_{n1} + V_{n,out}R_S/R_1$. Equating g_mV_{GS1} to $-V_{n,out}/R_1$ yields

$$g_m\left(\alpha V_{n,out} + V_{n1} + \frac{R_S}{R_1}V_{n,out}\right) = -\frac{V_{n,out}}{R_1}, \tag{5.117}$$

and hence

$$V_{n,out}|_{M1} = \frac{-g_m V_{n1}}{g_m(\alpha + \frac{R_S}{R_1}) + \frac{1}{R_1}}. \tag{5.118}$$

The noise current of R_1 is multiplied by the output impedance of the circuit, R_{out}. The reader can show that R_{out} is equal to R_1 in parallel with $(1 + g_m R_S)/(\alpha g_m)$. Summing this noise and that of M_1, dividing the result by the square of (5.116) and $4kTR_S$, and assuming the input is matched, we have

$$\text{NF} = 1 + \frac{\gamma}{g_m R_S} + \frac{R_S}{R_1}\left(1 + \frac{1}{g_m R_S}\right)^2. \tag{5.119}$$

That is, the NF can be lowered by raising g_m. Note that this result is identical to that expressed by Eq. (5.57) for the simple CG stage, except that $g_m R_S$ need not be equal to unity here. For example, if $g_m R_S = 4$ and $\gamma = 1$, then the first two terms yield a noise figure of 0.97 dB. In Problem 5.15 we reexamine these results if channel-length modulation is not neglected.

Example 5.19

How is the feedback factor, α, chosen in the above circuit?

Solution:

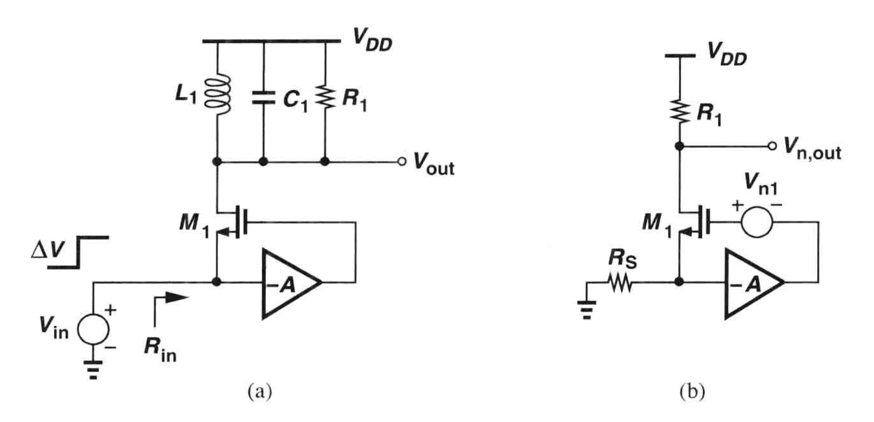
The design begins with the choice of $g_m R_S$ and $R_1/(2R_S)$ to obtain the required noise figure and voltage gain, A_v. For input matching, $g_m R_S - 1 = \alpha g_m R_1 = \alpha g_m (2 A_v R_S)$. It follows that

$$\alpha = \frac{g_m R_S - 1}{2 g_m R_S A_v}. \tag{5.120}$$

For example, if $g_m R_S = 4$ and $A_v = 6$ ($= 15.6$ dB), then $R_1 = 600\ \Omega$ and $\alpha = 1/16$.

Another variant of the CG LNA employs *feedforward* to avoid the tight relationship between the input resistance and the noise figure [4]. Illustrated in Fig. 5.45(a), the idea is to amplify the input by a factor of $-A$ and apply the result to the gate of M_1. For an input voltage change of ΔV, the gate-source voltage changes by $-(1 + A)\Delta V$ and the drain current by $-(1 + A)g_m \Delta V$. Thus, the g_m is "boosted" by a factor of $1 + A$ [4], *lowering* the input impedance to $R_{in} = [g_m(1 + A)]^{-1}$ and raising the voltage gain from the source to the drain to $(1 + A)g_m R_1$ (at resonance).

We now compute the noise figure with the aid of the equivalent circuit shown in Fig. 5.45(b). Since the current flowing through R_S is equal to $-V_{n,out}/R_1$, the source voltage is given by $-V_{n,out}R_S/R_1$ and the gate voltage by $(-V_{n,out}R_S/R_1)(-A) + V_{n1}$. Multiplying



(a) (b)

Figure 5.45 *(a) CG stage with feedforward, (b) calculation of NF.*

the gate-source voltage by g_m and equating the result to $-V_{n,out}/R_1$, we have

$$g_m\left(A\frac{R_S}{R_1}V_{n,out} + V_{n1} + \frac{R_S}{R_1}V_{n,out}\right) = -\frac{V_{n,out}}{R_1}, \tag{5.121}$$

and hence

$$V_{n,out}|_{M1} = \frac{-g_m R_1 V_{n1}}{(1+A)g_m R_S + 1}. \tag{5.122}$$

This expression reduces to $-g_m R_1 V_{n1}/2$ if the input is matched, indicating that half of the noise current of M_1 flows through R_1.[15] With input matching, the voltage gain from the left terminal of R_S in Fig. 5.45(b) to the output is equal to $(1 + A)g_m R_1/2$. We therefore sum the output noise contribution of M_1 and R_1, divide the result by the square of this gain and the noise of R_S, and add unity:

$$\text{NF} = 1 + \frac{\gamma}{1+A} + \frac{4R_S}{R_1}. \tag{5.123}$$

This equation reveals that the NF can be lowered by raising A with the constraint $g_m(1 + A) = R_S^{-1}$ (for input matching).

The above analysis has neglected the noise of the gain stage A in Fig. 5.45(a). We show in Problem 5.17 that the input-referred noise of this stage, V_{nA}^2, is multiplied by A and added to V_{n1} in Eq. (5.122), leading to an overall noise figure equal to

$$\text{NF} = 1 + \frac{\gamma}{1+A} + \frac{4R_S}{R_1} + \frac{A^2}{(1+A)^2}\frac{\overline{V_{nA}^2}}{4kTR_S}. \tag{5.124}$$

In other words, $\overline{V_{nA}^2}$ is referred to the input by a factor of $A^2/(1 + A)^2$, which is not much less than unity. For this reason, it is difficult to realize A by an active circuit.

It is possible to obtain the voltage gain through the use of an on-chip transformer. As shown in Fig. 5.46 [4], for a coupling factor of k between the primary and the secondary and

15. Where does the other half go?

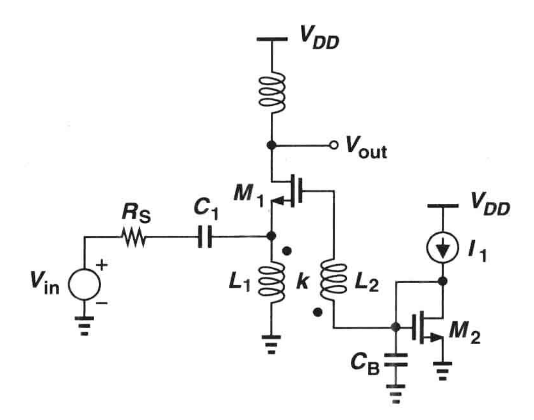


Figure 5.46 *CG stage with transformer feedforward.*

a turns ratio of n ($= \sqrt{L_2/L_1}$), the transformer provides a voltage gain of kn. The direction of the currents is chosen so as to yield a negative sign. However, on-chip transformer geometries make it difficult to achieve a voltage gain higher than roughly 3, even with stacked spirals [5]. Also, the loss in the primary and secondary contributes noise.

5.3.6 Noise-Cancelling LNAs

In our previous derivations of the noise figure of LNAs, we have observed three terms: a value of unity arising from the noise of R_S itself, a term representing the contribution of the input transistor, and another related to the noise of the load resistor. "Noise-cancelling LNAs" aim to cancel the second term [6]. The underlying principle is to identify *two* nodes in the circuit at which the signal appears with opposite polarities but the noise of the input transistor appears with the same polarity. As shown in Fig. 5.47, if nodes X and Y satisfy this condition, then their voltages can be properly scaled and summed such that the signal components add and the noise components cancel.

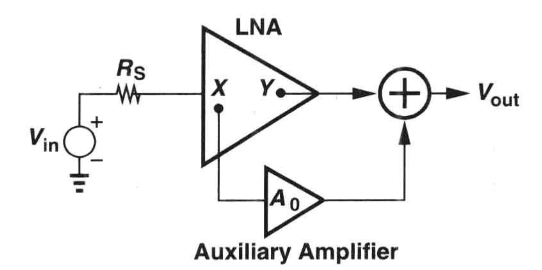


Figure 5.47 *Conceptual illustration of noise-cancelling LNAs.*

The CS stage with resistive feedback studied in Section 5.3.2 serves as a good candidate for noise cancellation because, as shown in Fig. 5.48(a), the noise current of M_1 flows through R_F and R_S, producing voltages at the gate and drain of the transistor with the same polarity. The signal, on the other hand, experiences inversion. Thus, as conceptually shown in Fig. 5.48(b), if V_X is amplified by $-A_1$ and added to V_Y, the noise of M_1 can be removed [6]. Since the noise voltages at nodes Y and X bear a ratio of $1 + R_F/R_S$ (why?), we choose $A_1 = 1 + R_F/R_S$. The signal experiences two additive gains: the original

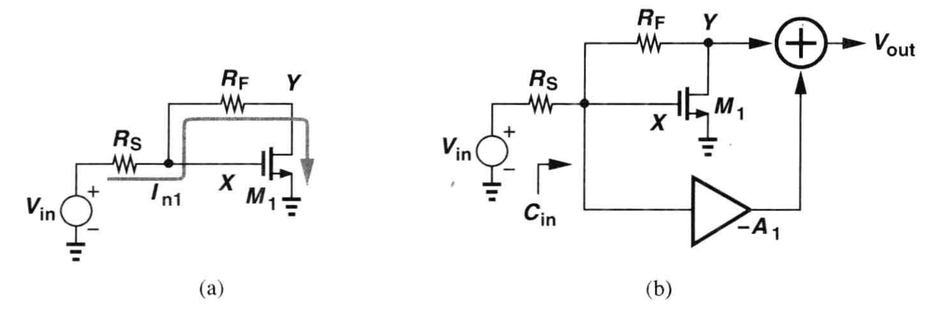


Figure 5.48 *(a) Noise of input transistor in a feedback CS stage, (b) cancellation of noise of M_1.*

gain, $V_Y/V_X = 1 - g_m R_F = 1 - R_F/R_S$ (if the input is matched), and the additional gain, $-(1 + R_F/R_S)$. It follows that

$$\frac{V_{out}}{V_X} = 1 - \frac{R_F}{R_S} - \left(1 + \frac{R_F}{R_S}\right) \tag{5.125}$$

$$= -\frac{2R_F}{R_S}, \tag{5.126}$$

if the input is matched. The gain V_{out}/V_{in} is half of this value.

Let us now compute the noise figure of the circuit, assuming that the auxiliary amplifier exhibits an input-referred noise voltage V_{nA1} and a high input impedance. Recall from Section 5.3.2 that the noise voltage of R_F appears directly at the output as $4kTR_F$. Adding this noise to $A_1^2\overline{V_{nA1}^2}$, dividing the result by $(R_F/R_S)^2$ and $4kTR_S$, and adding unity, we obtain the noise figure as

$$\text{NF} = 1 + \frac{R_S}{R_F} + A_1^2\overline{V_{nA1}^2}\frac{R_S}{4kTR_F^2}. \tag{5.127}$$

Since $A_1 = 1 + R_F/R_S$,

$$\text{NF} = 1 + \frac{R_S}{R_F} + \frac{\overline{V_{nA1}^2}}{4kTR_S}\left(1 + \frac{R_S}{R_F}\right)^2. \tag{5.128}$$

The NF can therefore be minimized by maximizing R_F and minimizing $\overline{V_{nA1}^2}$. Note that R_S/R_F is the inverse of the gain and hence substantially less than unity, making the third term approximately equal to $\overline{V_{nA1}^2}/(4kTR_S)$. That is, the noise of the auxiliary amplifier is directly referred to the input and must therefore be much less than that of R_S.

The input capacitance, C_{in}, arising from M_1 and the auxiliary amplifier degrades both S_{11} and the noise cancellation, thereby requiring a series (or parallel) inductor at the input for operation at very high frequencies. It can be proved [6] that the frequency-dependent noise figure is expressed as

$$\text{NF}(f) = \text{NF}(0) + [\text{NF}(0) - 1 + \gamma]\left(\frac{f}{f_0}\right)^2, \tag{5.129}$$

where NF(0) is given by (5.128) and $f_0 = 1/(\pi R_S C_{in})$.

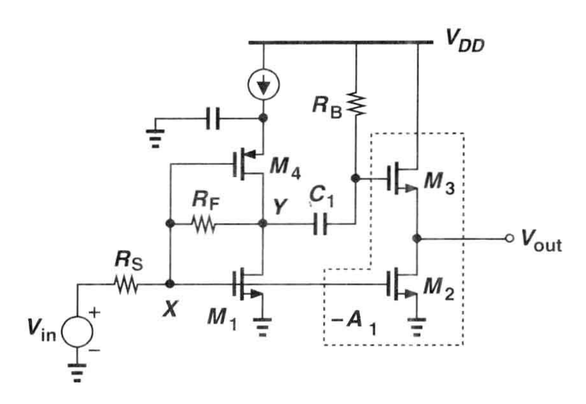


Figure 5.49 *Example of noise-cancelling LNA.*

Figure 5.49 depicts an implementation of the circuit [6]. Here, M_2 and M_3 serve as a CS amplifier, providing a voltage gain of $g_{m2}/(g_{m3} + g_{mb3})$, and also as the summing circuit. Transistor M_3 operates as a source follower, sensing the signal and noise at the drain of M_1. The first stage is similar to that studied in Example 5.7.

Example 5.20

Figure 5.50 shows an alternative implementation of a noise-cancelling LNA that also performs single-ended to differential conversion. Neglecting channel-length modulation, determine the condition for noise cancellation and derive the noise figure.

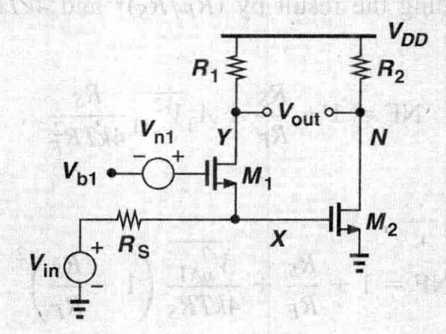


Figure 5.50 *CG/CS stage as a noise-cancelling LNA.*

Solution:

The circuit follows the noise cancellation principle because (a) the noise of M_1, V_{n1}, sees a source follower path to node X and a common-source path to node Y, exhibiting opposite polarities at these two nodes, and (b) the signal sees a common-gate path through X and Y, exhibiting the same polarity. Transistor M_1 produces half of its noise voltage at X if the input is matched (why?). Transistor M_2 senses this noise and amplifies it by a factor of $-g_{m2}R_2$. The reader can prove that the output noise of the CG stage due to M_1 (at Y) is equal to $(V_{n1}/2)g_{m1}R_1$. For noise cancellation, we must have

$$g_{m1}R_1\frac{V_{n1}}{2} = g_{m2}R_2\frac{V_{n1}}{2}, \tag{5.130}$$

Example 5.20 *(Continued)*

and, since $g_{m1} = 1/R_S$,

$$R_1 = g_{m2} R_2 R_S. \tag{5.131}$$

If the noise of M_1 is cancelled, the noise figure arises from the contributions of M_2, R_1, and R_2. The noise at Y is equal to $4kTR_1$ and at N equal to $4kT\gamma g_{m2} R_2^2 + 4kTR_2$. Since the total voltage gain, V_{out}/V_{in}, is given by $(g_{m1}R_1 + g_{m2}R_2)/2 = g_{m1}R_1 = R_1/R_S$, we have

$$\mathrm{NF} = 1 + \left(\frac{R_S}{R_1}\right)^2 (4kTR_1 + 4kT\gamma g_{m2} R_2^2 + 4kTR_2)\frac{1}{4kTR_S} \tag{5.132}$$

$$= 1 + \frac{R_S}{R_1} + \gamma\frac{R_2}{R_1} + \frac{R_S R_2}{R_1^2}. \tag{5.133}$$

The principal advantage of the above noise cancellation technique is that it affords the broadband characteristics of feedback or CG stages but with a lower noise figure. It is therefore suited to systems operating in different frequency bands or across a wide frequency range, e.g., 900 MHz to 5 GHz.

5.3.7 Reactance-Cancelling LNAs

It is possible to devise an LNA topology that inherently cancels the effect of its own input capacitance. Illustrated in Fig. 5.51(a) [7], the idea is to exploit the inductive input impedance of a negative-feedback amplifier so as to cancel the input capacitance, C_{in}. If the open-loop transfer function of the core amplifier is modeled by a one-pole response, $A_0/(1 + s/\omega_0)$, then the input admittance is given by

$$Y_1(s) = \frac{s + (A_0 + 1)\omega_0}{R_F(s + \omega_0)}. \tag{5.134}$$

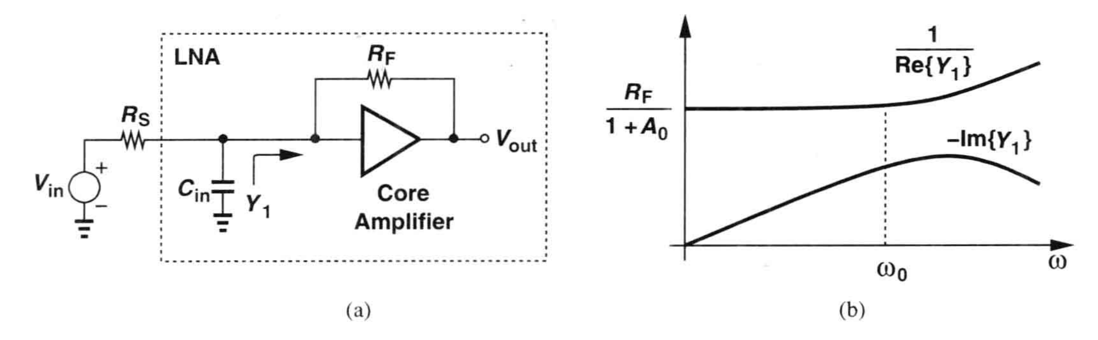


Figure 5.51 *(a) Reactance-cancelling LNA topology, (b) behavior of components of Y_1 with frequency.*

It follows that

$$\frac{1}{Re\{Y_1\}} = \frac{R_F(\omega^2 + \omega_0^2)}{(1 + A_0)\omega_0^2} \tag{5.135}$$

$$Im\{Y_1\} = \frac{-A_0\omega\omega_0}{R_F(\omega^2 + \omega_0^2)}. \tag{5.136}$$

At frequencies well below ω_0, $1/Re\{Y_1\}$ reduces to $R_F/(1 + A_0)$, which can be set equal to R_S, and $Im\{Y_1\}$ is roughly $-A_0\omega/(R_F\omega_0)$, which can be chosen to cancel $C_{in}\omega$. Figure 5.51(b) illustrates the behavior of $1/Re\{Y_1\}$ and $-Im\{Y_1\}$.

The input matching afforded by the above technique holds for frequencies up to about ω_0, dictating that the open-loop bandwidth of the core amplifier reach the maximum frequency of interest. The intrinsic speed of deep-submicron devices provides the gain and bandwidth required here.

The reader may wonder if our modeling of the core amplifier by a one-pole response applies to multistage implementations as well. We return to this point below.

Figure 5.52 shows a circuit realization of the amplifier concept for the frequency range of 50 MHz to 10 GHz [7]. Three common-source stages provide gain and allow negative feedback. Cascodes and source followers are avoided to save voltage headroom. The input transistor, M_1, has a large width commensurate with flicker noise requirements at 50 MHz, thus operating with a V_{GS} of about 200 mV. If this voltage also appears at node Y, it leaves no headroom for output swings, limiting the linearity of the circuit. To resolve this issue, current I_1 is drawn from R_F so as to shift up the quiescent voltage at Y by approximately 250 mV. Since $R_F = 1\ \mathrm{k}\Omega$, I_1 need be only 200 μA, contributing negligible noise at the LNA input.[16]

With three gain stages, the LNA can potentially suffer from a small phase margin and exhibit substantial peaking in its frequency response. In this design, the open-loop poles at nodes A, B, X, and Y lie at 10 GHz, 24.5 GHz, 22 GHz, and 75 GHz, respectively, creating a great deal of phase shift. Nonetheless, due to the small feedback factor,

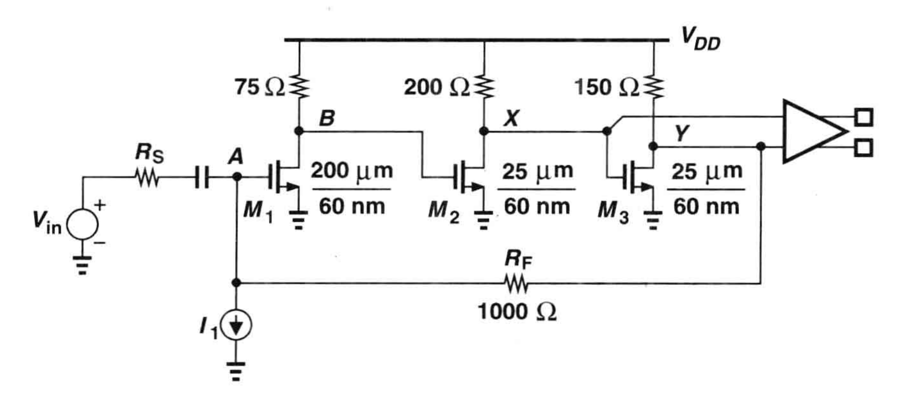


Figure 5.52 *Implementation of reactance-cancelling LNA.*

16. Alternatively, capacitive coupling can be used in the feedback path. But the large value necessary for the capacitor would introduce additional parasitics.

$R_S/(R_S + R_F) = 0.048$, simulations indicate that the circuit provides a phase margin of about 50° and a peaking of 1 dB in its closed-loop frequency response.

The multi-pole LNA of Fig. 5.52 contains an inductive component in its input impedance but with a behavior more complex than the above analysis suggests. Fortunately, behavioral simulations confirm that, if the poles at B, X, and Y are "lumped" (i.e., their time constants are added), then the one-pole approximation still predicts the input admittance accurately. The pole frequencies mentioned above collapse to an equivalent value of $\omega_0 = 2\pi(9.9\text{ GHz})$, suggesting that the real and imaginary parts of Y_1 retain the desired behavior up to the edge of the cognitive radio band.

The LNA output is sensed between nodes X and Y. Even though these nodes provide somewhat unequal swings and a phase difference slightly greater than 180°, the pseudo-differential sensing still raises both the gain and the IP_2, the latter because second-order distortion at X also appears at Y and is thus partially cancelled in $V_Y - V_X$.[17]

5.4 GAIN SWITCHING

The dynamic range of the signal sensed by a receiver may approach 100 dB. For example, a cell phone may receive a signal level as high as −10 dBm if it is close to a base station or as low as −110 dBm if it is in an underground garage. While designed for the highest sensitivity, the receiver chain must still detect the signal correctly as the input level continues to increase. This requires that the gain of each stage be reduced so that the subsequent stages remain sufficiently linear with the large input signal. Of course, as the gain of the receiver is reduced, its noise figure rises. The gain must therefore be lowered such that the degradation in the sensitivity is less than the increase in the received signal level, i.e., the SNR does not fall. Figure 5.53 shows a typical scenario.

Gain switching in an LNA must deal with several issues: (1) it must negligibly affect the input matching; (2) it must provide sufficiently small "gain steps"; (3) the additional devices performing the gain switching must not degrade the speed of the original LNA;

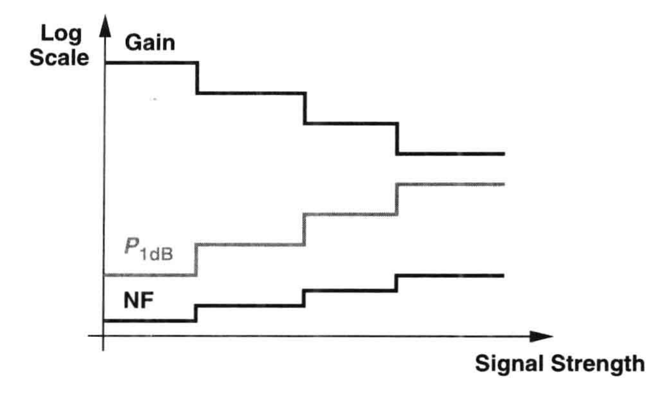


Figure 5.53 *Effect of gain switching on NF and P_{1dB}.*

17. To ensure stability in the presence of package parasitics, a capacitor of 10-20 pF must be placed between V_{DD} and GND.

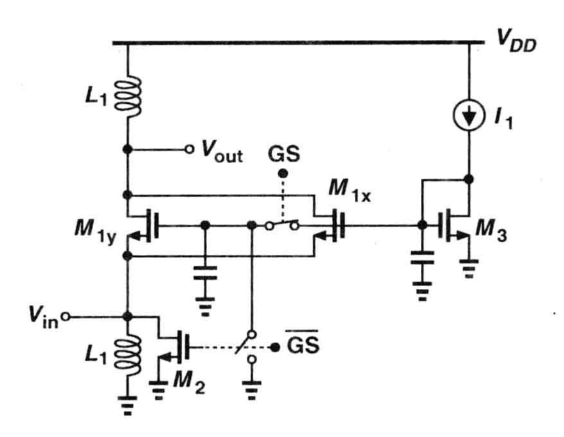


Figure 5.54 *Example of gain switching in CG stage.*

(4) for high input signal levels, gain switching must also make the LNA *more linear* so that this stage does not limit the receiver linearity. As seen below, some LNA topologies lend themselves more easily to gain switching than others do.

Let us first consider a common-gate stage. Can we reduce the transconductance of the input transistor to reduce the gain? To switch the gain while maintaining input matching, we can insert a physical resistance in parallel with the input as g_m is lowered. Figure 5.54 shows an example [8], where the input transistor is decomposed into two, M_{1x} and M_{1y}, and transistor M_2 introduces a parallel resistance if it is on. In the "high-gain mode," the gain select line, GS, is high, placing M_{1x} and M_{1y} in parallel, and M_2 is off. In the "low-gain mode," M_{1y} turns off, reducing the gain, and M_2 turns on, ensuring that $R_{on2}||(g_{m1x} + g_{mb1x})^{-1} = R_S$. For example, to reduce the gain by 6 dB, we choose equal dimensions for M_{1x} and M_{1y} and $R_{on2} = (g_{m1x} + g_{mb1x})^{-1} = 2R_S$ (why?). Also, the gate of M_{1y} is secured to ground by a capacitor to avoid the on-resistance of the switch at high frequencies.

Example 5.21

Choose the devices in the above circuit for a gain step of 3 dB.

Solution:

To reduce the voltage gain by $\sqrt{2}$, we have

$$\frac{W_{1x}}{W_{1x} + W_{1y}} = \frac{1}{\sqrt{2}}, \tag{5.137}$$

and hence $W_{1y}/W_{1x} = \sqrt{2} - 1$. We also note that, with M_{1y} off, the input resistance rises to $\sqrt{2}R_S$. Thus, $R_{on2}||(\sqrt{2}R_S) = R_S$ and hence

$$R_{on2} = \frac{\sqrt{2}}{\sqrt{2} - 1}R_S. \tag{5.138}$$

In Problem 5.20, we calculate the noise figure after the 3-dB gain reduction.

In the above calculation, we have neglected the effect of channel-length modulation. If the upper bound expressed by Eq. (5.67) restricts the design, then the cascode CG stage of Fig. 5.24 can be used.

Another approach to switching the gain of a CG stage is illustrated in Fig. 5.55, where the on-resistance of M_2 appears in parallel with R_1. With input matching and in the absence of channel-length modulation, the gain is given by

$$\frac{V_{out}}{V_{in}} = \frac{R_1||R_{on2}}{2R_S}. \tag{5.139}$$

For multiple gain steps, a number of PMOS switches can be placed in parallel with R_1. The following example elaborates on this point.

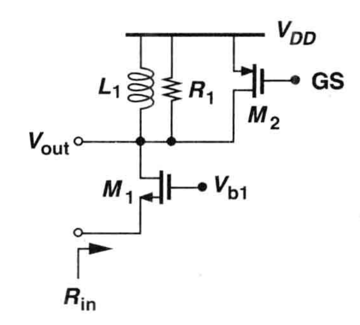


Figure 5.55 *Effect of load switching on input impedance.*

Example 5.22

Design the load switching network of Fig. 5.55 for two 3-dB gain steps.

Solution:

As shown in Fig. 5.56, M_{2a} and M_{2b} switch the gain. For the first 3-dB reduction in gain, M_{2a} is turned on and

$$R_1||R_{on,a} = \frac{R_1}{\sqrt{2}}, \tag{5.140}$$

Figure 5.56 *Load switching for 3-dB gain steps.*

(Continues)

Example 5.22 *(Continued)*

i.e., $R_{on,a} = R_1/(\sqrt{2}-1)$. For the second 3-dB reduction, *both* M_{2a} and M_{2b} are turned on and

$$R_1 || R_{on,a} || R_{on,b} = \frac{R_1}{2}, \tag{5.141}$$

i.e., $R_{on,b} = R_1/(2-\sqrt{2})$. Note that if only M_{2b} were on in this case, then it would need to be wider, thus contributing a greater capacitance to the output node.

The principal difficulty with switching the load resistance in a CG stage is that it alters the input resistance, as expressed by $R_{in} = (R_1 + r_O)/(1 + g_m r_O)$. This effect can be minimized by adding a cascode transistor as in Fig. 5.24. The use of a cascode transistor also permits a third method of gain switching. Illustrated in Fig. 5.57, the idea is to route part of the drain current of the input device to V_{DD}—rather than to the load—by means of another cascode transistor, M_3. For example, if M_2 and M_3 are identical, then turning M_3 on yields $\alpha = 0.5$, dropping the voltage gain by 6 dB.

The advantage of the above technique over the previous two is that the gain step depends only on W_3/W_2 (if M_2 and M_3 have equal lengths) and not the absolute value of the on-resistance of a MOS switch. The bias and signal currents produced by M_1 split between M_3 and M_2 in proportion to W_3/W_2, yielding a gain change by a factor of $1 + W_3/W_2$. As a result, gain steps in the circuit of Fig. 5.57 are more accurate than those in Figs. 5.54 and 5.55. However, the capacitance introduced by M_3 at node Y degrades the performance at high frequencies. For a single gain step of 6 dB, we have $W_3 = W_2$, nearly doubling the capacitance at this node. For a gain reduction by a factor of N, $W_3 = (N-1)W_2$, possibly degrading the performance considerably.

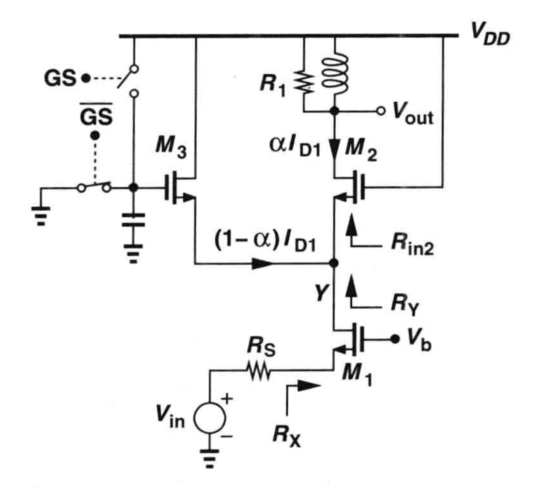


Figure 5.57 *Gain switching by cascode device.*

Example 5.23

If $W_3 = W_2$ in Fig. 5.57, how does the input impedance of the circuit change from the high-gain mode to the low-gain mode? Neglect body effect.

Solution:

In the high-gain mode, the input impedance is given by Eq. (5.70). In the low-gain mode, the impedance seen looking into the source of M_2 changes because both g_{m2} and r_{O2} change. For a square-law device, a twofold reduction in the bias current (while the dimensions remain unchanged) translates to a twofold increase in r_O and a $\sqrt{2}$ reduction in g_m. Thus, in Fig. 5.57,

$$R_{in2} = \frac{R_1 + 2r_{O2}}{1 + \sqrt{2}g_{m2}r_{O2}}, \tag{5.142}$$

where g_{m2} and r_{O2} correspond to the values while M_3 is off. Transistor M_3 presents an impedance of $(1/g_{m3})||r_{O3}$ at Y, yielding

$$R_Y = \frac{1}{g_{m3}}||r_{O3}||\frac{R_1 + 2r_{O2}}{1 + \sqrt{2}g_{m2}r_{O2}}. \tag{5.143}$$

Transistor M_1 transforms this impedance to

$$R_X = \frac{R_Y + r_{O1}}{1 + g_{m1}r_{O1}}. \tag{5.144}$$

This impedance is relatively independent of the gain setting because R_Y is on the order of $1/g_m$.

In order to reduce the capacitance contributed by the gain switching transistor, we can *turn off* part of the main cascode transistor so as to create a greater imbalance between the two. Shown in Fig. 5.58 (on page 310) is an example where M_2 is decomposed into two devices so that, when M_3 is turned on, M_{2a} is turned off. Consequently, the gain drops by a factor of $1 + W_3/W_{2b}$ rather than $1 + W_3/(W_{2b} + W_{2a})$.

Example 5.24

Design the gain switching network of Fig. 5.58 for two 3-dB steps. Assume equal lengths for the cascode devices.

Solution:

To reduce the gain by 3 dB, we turn on M_3 while M_{2a} and M_{2b} remain on. Thus,

$$1 + \frac{W_3}{W_{2a} + W_{2b}} = \sqrt{2}. \tag{5.145}$$

(*Continues*)

Example 5.24 *(Continued)*

For another 3-dB reduction, we turn off M_{2b}:

$$1 + \frac{W_3}{W_{2a}} = 2. \tag{5.146}$$

It follows from Eqs. (5.145) and (5.146) that

$$W_3 = W_{2a} = \frac{W_{2b}}{\sqrt{2}}. \tag{5.147}$$

In a more aggressive design, M_2 would be decomposed into *three* devices, such that one is turned off for the first 3-dB step, allowing M_3 to be narrower. The calculations are left as an exercise for the reader.

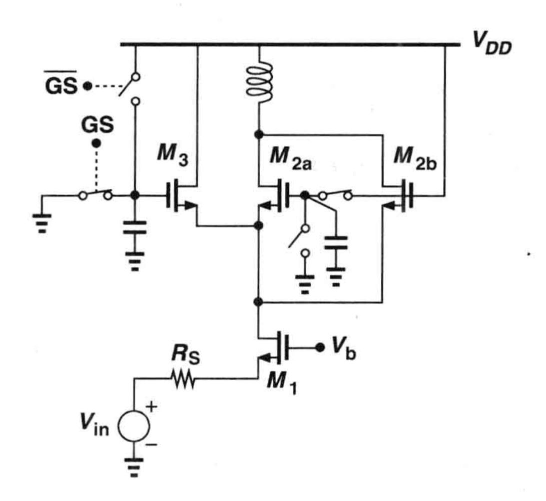


Figure 5.58 *Gain switching by programmable cascode devices.*

We now turn our attention to gain switching in an inductively-degenerated cascode LNA. Can we switch part of the input transistor to switch the gain (Fig. 5.59)? Turning M_{1b} off does not alter ω_T because the current density remains constant. Thus, $Re\{Z_{in}\} = L_1\omega_T$ is relatively constant, but $Im\{Z_{in}\}$ changes, degrading the input match. If the input match is somehow restored, then the voltage gain, $R_1/(2L_1\omega)$, does not change! Furthermore, the thermal noise of S_1 degrades the noise figure in the high gain mode. For these reasons, gain switching must be realized in other parts of the circuit.

As with the CG LNA of Fig. 5.55, the gain can be reduced by placing one or more PMOS switches in parallel with the load [Fig. 5.60(a)]. Alternatively, the cascode switching scheme of Fig. 5.57 can be applied here as well [Fig. 5.60(b)]. The latter follows the calculations outlined in Example 5.24, providing well-defined gain steps with a moderate

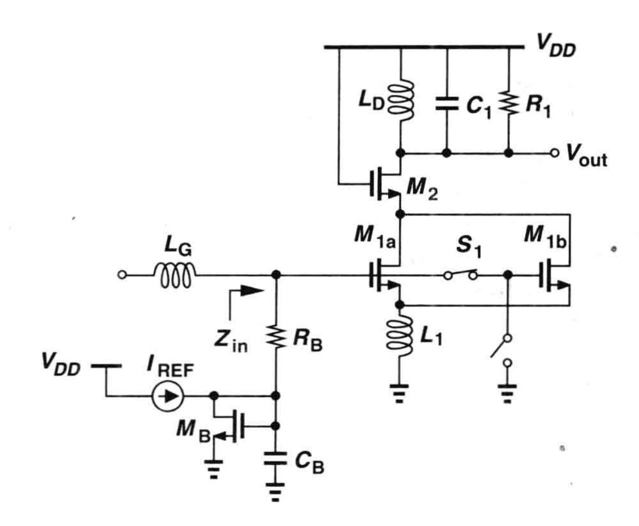


Figure 5.59 *Gain switching in CS stage.*

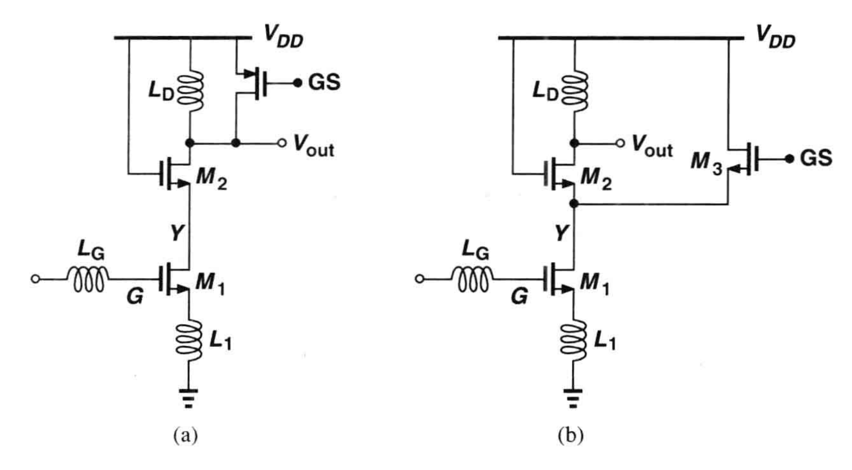


Figure 5.60 *Gain switching in cascode CS stage by (a) load switching, (b) additional cascode device.*

additional capacitance at node Y. It is important to bear in mind that cascode switching is attractive because it reduces the current flowing through the load by a well-defined ratio *and* it negligibly alters the input impedance of the LNA.

For the two variants of the CG stage studied in Section 5.3.3, gain switching can be realized by cascode devices as illustrated in Fig. 5.57. The use of feedback or feedforward in these topologies makes it difficult to change the gain through the input transistor without affecting the input match.

Lastly, let us consider gain switching in the noise-cancelling LNA of Fig. 5.48(b). Since $V_Y/V_X = 1 - R_F/R_S$ and R_{in} is approximately equal to $1/g_{m1}$ and independent of R_F, the gain can be reduced simply by lowering the value of R_F. Though not essential in the low-gain mode, noise cancellation can be preserved by adjusting A_1 so that it remains equal to $1 + R_F/R_S$.

Which one of the foregoing gain reduction techniques also makes the LNA *more linear*? None, except for the last one! Since the CG and CS stages retain the gate-source voltage swing (equal to half of the input voltage swing), their linearity improves negligibly. In the feedback LNA of Fig. 5.48(b), on the other hand, a lower R_F strengthens the negative feedback, raising the linearity to some extent.

Receiver designs in which the LNA nonlinearity becomes problematic at high input levels can "bypass" the LNA in very-low-gain modes. Illustrated conceptually in Fig. 5.61, the idea is to omit the LNA from the signal path so that the mixer (presumably more linear) directly senses the received signal. The implementation is not straightforward if input matching must be maintained. Figure 5.62 depicts a common-gate example, where M_1 is turned off, M_2 is turned on to produce a 50-Ω resistance, and M_3 is turned on to route the signal to the mixer.

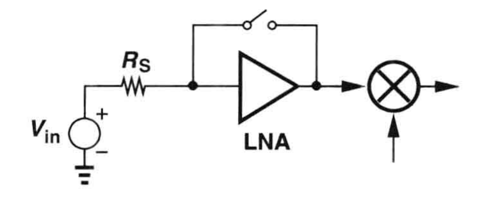


Figure 5.61 *LNA bypass.*

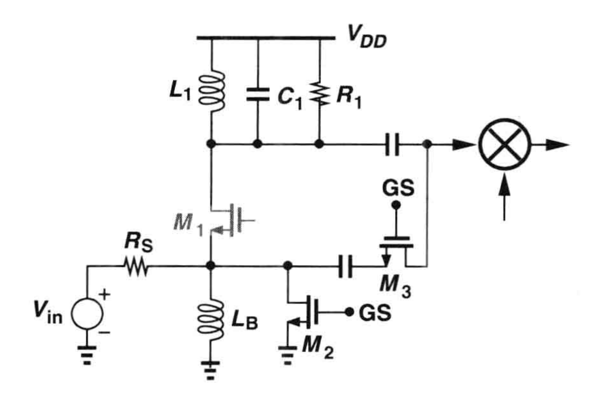


Figure 5.62 *Realization of LNA bypass.*

5.5 BAND SWITCHING

As mentioned in Section 5.1, LNAs that must operate across a wide bandwidth or in different bands can incorporate band switching. Figure 5.63(a) repeats the structure of Fig. 5.7(a), with the switch realized by a MOS transistor. Since the bias voltage at the output node is near V_{DD}, the switch must be a PMOS device, thus contributing a larger capacitance for a given on-resistance than an NMOS transistor. This capacitance lowers the tank resonance frequency when S_1 is *off*, reducing the maximum tolerable value of C_1 and hence limiting the size of the input transistor of the following stage. (If L_1 is reduced to compensate for the higher capacitance, then so are R_1 and the gain.) For this reason, we prefer the implementation in Fig. 5.63(b), where S_1 is formed as an NMOS device tied to ground.

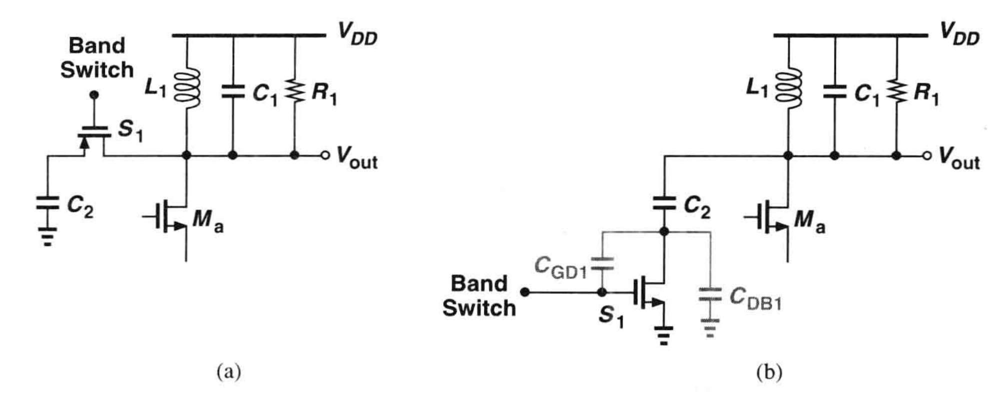


Figure 5.63 *(a) Band switching, (b) effect of switch parasitics.*

The choice of the width of S_1 in Fig. 5.63(b) proves critical. For a very narrow transistor, the on-resistance, R_{on1}, remains so high that the tank does not "feel" the presence of C_2 when S_1 is on. For a moderate device width, R_{on1} limits the Q of C_2, thereby lowering the Q of the overall tank and hence the voltage gain of the LNA. This can be readily seen by transforming the series combination of C_2 and R_{on1} to a parallel network consisting of C_2 and $R_{P1} \approx Q^2 R_{on1}$, where $Q = (C_2 \omega R_{on1})^{-1}$. That is, R_1 is now shunted by a resistance $R_{P1} = (C_2^2 \omega^2 R_{on1})^{-1}$.

The foregoing observation implies that R_{on1} must be minimized such that $R_{P1} \gg R_1$. However, as the width of S_1 in Fig. 5.63(b) increases, so does the capacitance that it introduces in the *off state*. The equivalent capacitance seen by the tank when S_1 is off is equal to the series combination of C_2 and $C_{GD1} + C_{DB1}$, which means C_1 must be less than its original value by this amount. We therefore conclude that the width of S_1 poses a trade-off between the tolerable value of C_1 when S_1 is off and the reduction of the gain when S_1 is on. (Recall that C_1 arises from M_a, the input capacitance of the next stage, and the parasitic capacitance of L_1.)

An alternative method of band switching incorporates two or more tanks as shown in Fig. 5.64 [8]. To select one band, the corresponding cascode transistor is turned on while the other remains off. This scheme requires that each tank drive a copy of the following stage, e.g., a mixer. Thus, when M_1 and band 1 are activated, so is mixer MX_1. The principal drawback of this approach is the capacitance contributed by the additional cascode device(s) to node Y. Also, the spiral inductors have large footprints, making the layout and routing more difficult.

5.6 HIGH-IP$_2$ LNAS

As explained in Chapter 4, even-order distortion can significantly degrade the performance of direct-conversion receivers. Since the circuits following the downconversion mixers are typically realized in differential form,[18] they exhibit a high IP$_2$, leaving the LNA and the

18. And since they employ large devices and hence have small mismatches.

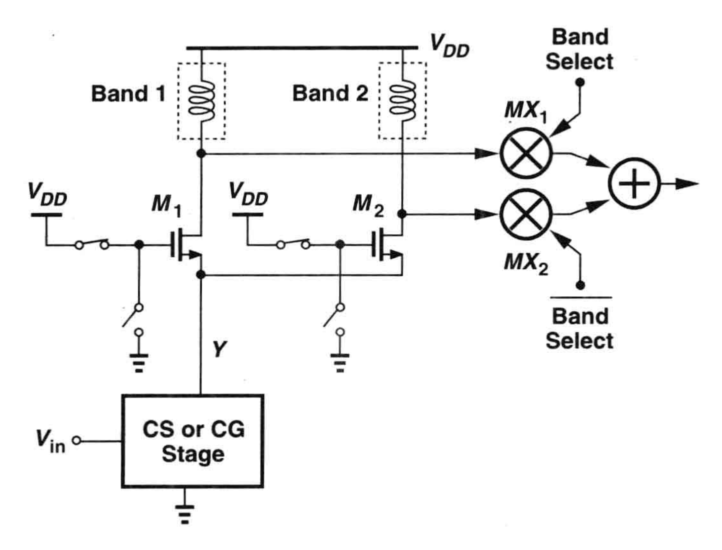


Figure 5.64 *Band switching by programmable cascode branches.*

mixers as the IP_2 bottleneck of the receivers. In this section, we study techniques of raising the IP_2 of LNAs, and in Chapter 6, we do the same for mixers.

5.6.1 Differential LNAs

Differential LNAs can achieve high IP_2's because, as explained in Chapter 2, symmetric circuits produce no even-order distortion. Of course, some (random) asymmetry plagues actual circuits, resulting in a finite, but still high, IP_2.

In principle, any of the single-ended LNAs studied thus far can be converted to differential form. Figure 5.65 depicts two examples. Not shown here, the bias network for the input transistors is similar to those described in Sections 5.3.3 and 5.3.4.

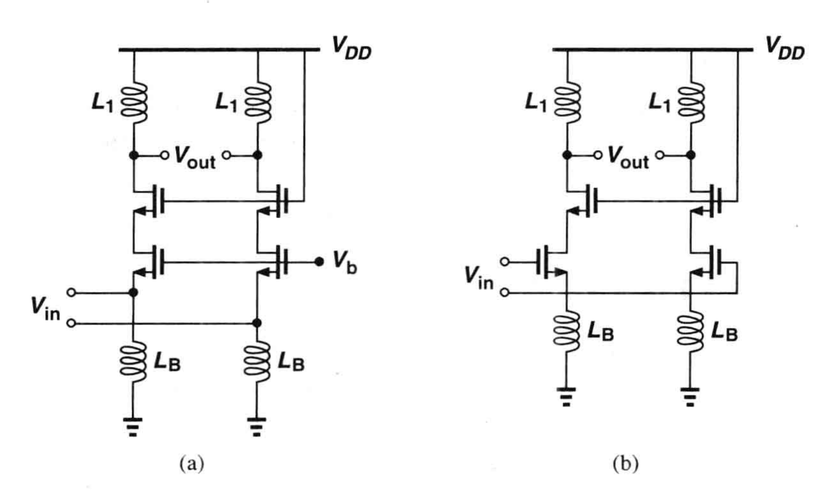


Figure 5.65 *Differential (a) CG and (b) CS stages.*

But what happens to the noise figure of the circuit if it is converted to differential form? Before answering this question, we must determine the source impedance driving the LNA. Since the antenna and the preselect filter are typically single-ended, a transformer must precede the LNA to perform single-ended to differential conversion. Illustrated in Fig. 5.66(a), such a cascade processes the signal differentially from the input port of the LNA to the end of the baseband section. The transformer is called a "balun," an acronym for "balanced-to-unbalanced" conversion because it can also perform differential to single-ended conversion if its two ports are swapped.

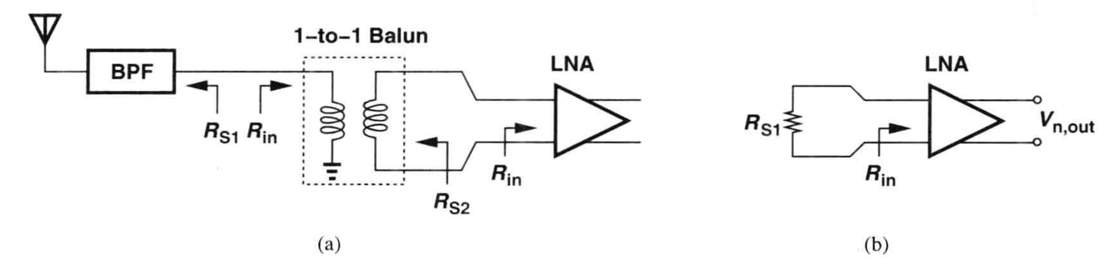


Figure 5.66 *(a) Use of balun at RX input, (b) simplified circuit.*

If the source impedance provided by the antenna and the band-pass filter in Fig. 5.66(a) is R_{S1} (e.g., 50 Ω), what is the differential source impedance seen by the LNA, R_{S2}? For a lossless 1-to-1 balun, i.e., for a lossless transformer with an equal number of turns in its primary and secondary, we have $R_{S2} = R_{S1}$. We must thus obtain the noise figure of the differential LNA with respect to a differential source impedance of R_{S1}. Figure 5.66(b) shows the setup for output noise calculation.

Note that the differential input impedance of the LNA, R_{in}, must be equal to R_{S1} for proper input matching. Thus, in the LNAs of Figs. 5.66(a) and (b), the *single-ended* input impedance of each half circuit must be equal to $R_{S1}/2$, e.g., 25 Ω.

Differential CG LNA We now calculate the noise figure of the differential CG LNA of Fig. 5.65(a), assuming it is designed such that the impedance seen between each input node and ground is equal to $R_{S1}/2$. In other words, each CG transistor must provide an input resistance of 25 Ω. Figure 5.67(a) shows the simplified environment, emphasizing that the noise figure is calculated with respect to a source impedance of R_{S1}. Redrawing Fig. 5.67(a) as shown in Fig. 5.67(b), we recognize from the symmetry of the circuit that

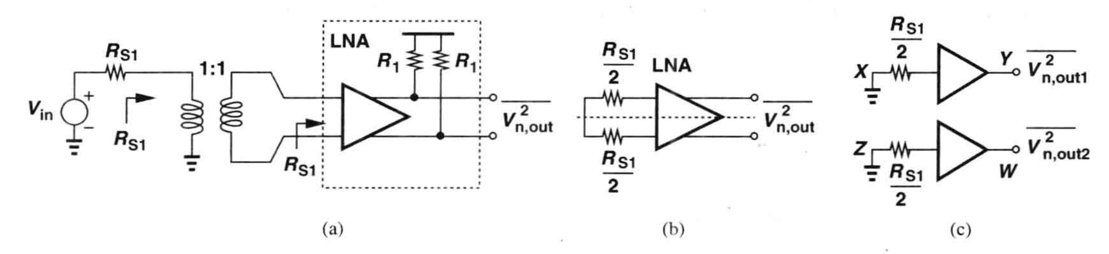


Figure 5.67 *(a) Cascade of balun and LNA, (b) simplified circuit of (a), and (c) simplified circuit of (b).*

we can compute the output noise of each half circuit as in Fig. 5.67(c) and add the output powers:

$$\overline{V_{n,out}^2} = \overline{V_{n,out1}^2} + \overline{V_{n,out2}^2}. \tag{5.148}$$

Since each half circuit provides matching at the input, the CG results of Section 5.3.3 apply here as well with the substitution $R_S = R_{S1}/2$. Specifically, the voltage gain from X to Y is equal to $R_1/(2R_{S1}/2)$, where R_1 denotes the load resistance of the CG half circuit. The output noise consists of (1) the input transistor contribution, given by Eq. (5.56), (2) the load resistor contribution, $4kTR_1$, and (3) the source impedance contribution, $(4kTR_{S1}/2)[R_1/(2R_1/2)]$:

$$\overline{V_{n,out1}^2} = kT\gamma\frac{R_1^2}{R_{S1}/2} + 4kTR_1 + 4kT\frac{R_{S1}}{2}\left(\frac{R_1}{\frac{2R_{S1}}{2}}\right)^2. \tag{5.149}$$

From Eq. (5.148), the total output noise power is twice this amount. Noting that the total voltage gain $A_v = (V_Y - V_W)/(V_X - V_Z)$ is equal to that of half of the circuit, V_Y/V_X $(= R_1/R_{S1})$, we compute the noise figure with respect to a source impedance of R_{S1} as

$$\text{NF} = \frac{\overline{V_{n,out}^2}}{A_v^2} \cdot \frac{1}{4kTR_{S1}} \tag{5.150}$$

$$= 1 + \gamma + \frac{2R_{S1}}{R_1}. \tag{5.151}$$

Interestingly, this value is lower than that of the single-ended counterpart [Eq. (5.58)]. But why? Since in Fig. 5.67(c), $V_Y/V_X = R_1/(2R_{S1}/2) = R_1/R_{S1}$, we observe that the voltage gain is twice that of the single-ended CG LNA. (After all, the transconductance of the input transistor is doubled to lower the input impedance to $R_{S1}/2$.) On the other hand, the overall differential circuit contains *two* R_1's at its output, each contributing a noise power of $4kTR_1$. The total, $8kTR_1$, divided by $(R_1/R_{S1})^2$ and $4kTR_{S1}$ yields $2R_{S1}/R_1$. Of course, the value stipulated by Eq. (5.151) can be readily obtained in a single-ended CG LNA by simply doubling the load resistance. Figure 5.68 summarizes the behavior of the two circuits, highlighting the greater voltage gain in the differential topology. If identical gains are desired, the value of the load resistors in the differential circuit must be halved, thereby yielding identical noise figures.

In summary, a single-ended CG LNA can be converted to differential form according to one of three scenarios: (1) simply copy the circuit, in which case the differential input resistance reaches 100 Ω, failing to provide matching with a 1-to-1 balun; (2) copy the circuit but double the transconductance of the input transistors, in which case the input is matched but the overall voltage gain is doubled; (3) follow the second scenario but halve the load resistance to retain the same voltage gain. The second choice is generally preferable. Note that, for a given noise figure, a differential CG LNA consumes *four* times the power of a single-ended stage.[19]

19. To halve the input resistance, the transistor width and bias current must be doubled.

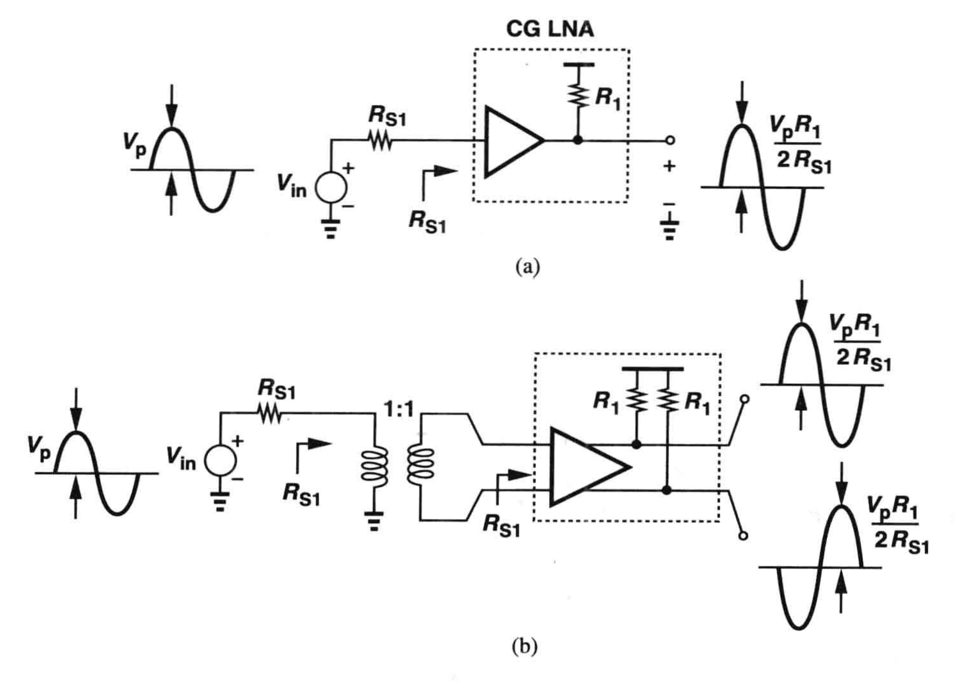


Figure 5.68 *Comparison of (a) single-ended and (b) differential CG LNAs.*

Our NF calculations have assumed an ideal balun. In reality, even external baluns have a loss as high as 0.5 dB, raising the NF by the same amount.

Example 5.25

An amplifier having a high input impedance employs a parallel resistor at the input to provide matching [Fig. 5.69(a)]. Determine the noise figure of the circuit and its differential version, shown in Fig. 5.69(b), where two replicas of the amplifier are used.

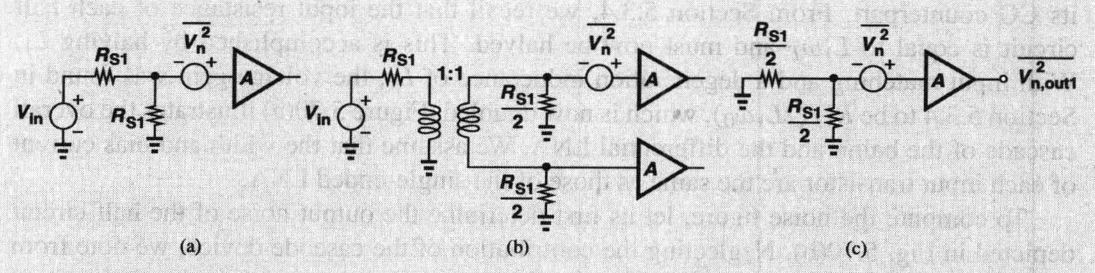


Figure 5.69 *(a) NF of an LNA with resistive termination, (b) differential version of (a), (c) simplified circuit of (b).*

Solution:

In the circuit of Fig. 5.69(a), the amplifier input-referred noise current is negligible and the total noise at the output is equal to $(4kTR_{S1}/2)A^2 + A^2\overline{V_n^2}$. The noise figure of the

(Continues)

Example 5.25 *(Continued)*

single-ended circuit is therefore given by

$$\mathrm{NF}_{sing} = \frac{4kT\dfrac{R_{S1}}{2}A^2 + A^2\overline{V_n^2}}{\dfrac{A^2}{4}} \cdot \frac{1}{4kTR_{S1}} \quad (5.152)$$

$$= 2 + \frac{\overline{V_n^2}}{kTR_{S1}}. \quad (5.153)$$

For the differential version, we write from the simplified half circuit shown in Fig. 5.69(c), $\overline{V_{n,out1}^2} = (4kTR_{S1}/4)A^2 + A^2\overline{V_n^2}$. The total output noise power of the differential circuit is twice this amount. The corresponding noise figure is then given by

$$\mathrm{NF}_{diff} = \frac{2\left(4kT\dfrac{R_{S1}}{4}A^2 + A^2\overline{V_n^2}\right)}{\dfrac{A^2}{4}} \cdot \frac{1}{4kTR_{S1}} \quad (5.154)$$

$$= 2 + \frac{2\overline{V_n^2}}{kTR_{S1}}. \quad (5.155)$$

In this case, the noise figure of the differential circuit is *higher*. We conclude that whether the differential version of an LNA exhibits a higher or lower NF depends on the circuit topology.

Differential CS LNA The differential CS LNA of Fig. 5.65(b) behaves differently from its CG counterpart. From Section 5.3.4, we recall that the input resistance of each half circuit is equal to $L_1\omega_T$ and must now be halved. This is accomplished by halving L_1. With input matching and a degeneration inductance of L_1, the voltage gain was found in Section 5.3.4 to be $R_1/(2L_1\omega_0)$, which is now doubled. Figure 5.70(a) illustrates the overall cascade of the balun and the differential LNA. We assume that the width and bias current of each input transistor are the same as those of the single-ended LNA.

To compute the noise figure, let us first determine the output noise of the half circuit depicted in Fig. 5.70(b). Neglecting the contribution of the cascode device, we note from Section 5.3.4 that, if the input is matched, half of the noise current of the input transistor flows from the output node. Thus,

$$\overline{V_{n,out1}^2} = kT\gamma g_{m1}R_1^2 + 4kTR_1 + 4kT\frac{R_{S1}}{2}\left(\frac{R_1}{L_1\omega_0}\right)^2. \quad (5.156)$$

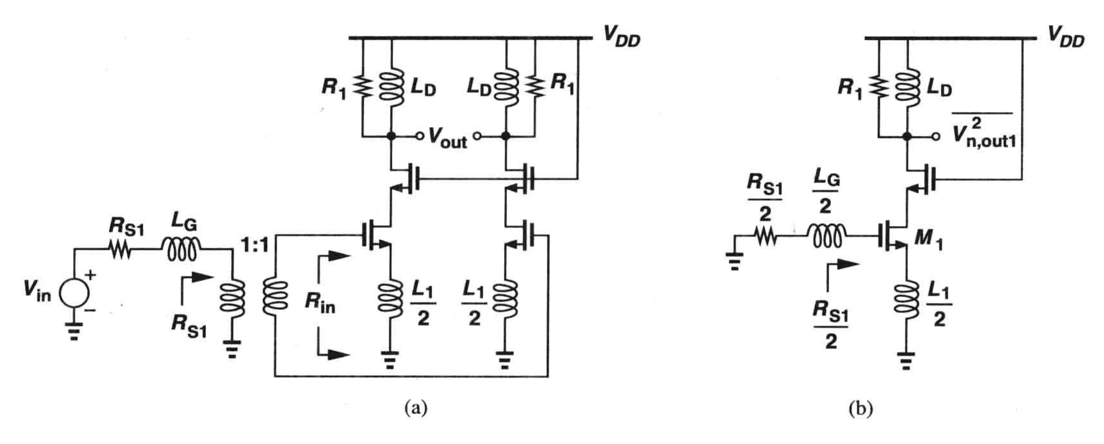


Figure 5.70 *(a) Differential CS LNA and (b) its half circuit.*

Multiplying this power by two, dividing it by $A_v^2 = R_1^2/(L_1\omega_0)^2$ and $4kTR_{S1}$, and noting that $L_1\omega_T/2 = R_{S1}/2$, we have

$$\text{NF} = \frac{\gamma}{2} g_{m1} R_{S1} \left(\frac{\omega_0}{\omega_T}\right)^2 + \frac{2R_{S1}}{R_1}\left(\frac{\omega_0}{\omega_T}\right)^2 + 1. \tag{5.157}$$

How does this compare with the noise figure of the original single-ended LNA [Eq. (5.101)]? We observe that both the transistor contribution and the load contribution are halved. The transistor contribution is halved because g_{m1} and hence the transistor noise *current* remain unchanged while the overall *transconductance* of the circuit is doubled. To understand this point, recall from Section 5.3.4 that $G_m = \omega_T/(2\omega_0 R_S)$ for the original single-ended circuit. Now consider the equivalent circuit shown in Fig. 5.71, where the differential transconductance, $(I_1 - I_2)/V_{in}$, is equal to $\omega_T/(\omega_0 R_{S1})$ (why?). The differential output current contains the noise currents of both M_1 and M_2 and is equal to $2(kT\gamma g_{m1})$. If this power is divided by the square of the transconductance and $4kTR_{S1}$, the first term in Eq. (5.157) is obtained.

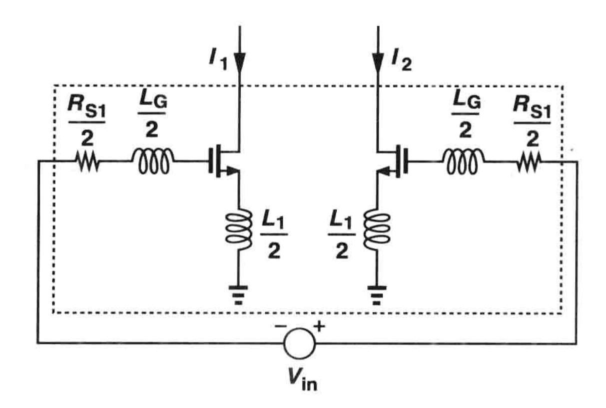


Figure 5.71 *Differential CS stage viewed as a transconductor.*

The reduction of the input transistor noise contribution in Eq. (5.157) is a remarkable property of differential operation, reinforcing the NF advantage of the degenerated CS stage over the CG LNA. However, this result holds only if the design can employ *two* degeneration inductors, each having *half* the value of that in the single-ended counterpart. This is difficult with bond wires as their physical length cannot be shortened arbitrarily. Alternatively, the design can incorporate on-chip degeneration inductors while converting the effect of the (inevitable) bond wire to a common-mode inductance. Figure 5.72 shows such a topology. With perfect symmetry, the bond wire inductance has no effect on the differential impedance seen between the gates. Nonetheless, as explained in Chapter 7, on-chip inductors suffer from a low quality factor (e.g., a high series resistance), possibly degrading the noise figure. We compare the power consumptions of the single-ended and differential implementations in Problem 5.21.

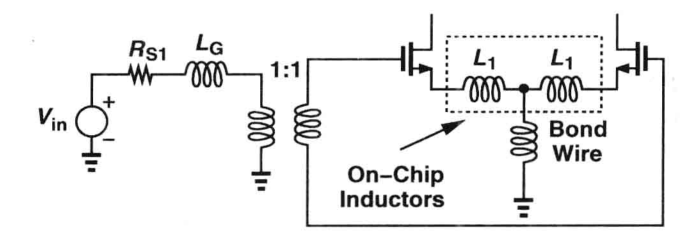


Figure 5.72 *Differential CS stage with on-chip degeneration inductors.*

The NF advantage implied by Eq. (5.157) may not materialize in reality because the loss of the balun is not negligible.

Is it possible to use a differential pair to convert the single-ended antenna signal to differential form? As shown in Fig. 5.73(a), the signal is applied to one input while the other is tied to a bias voltage. At low to moderate frequencies, V_X and V_Y are differential and the voltage gain is equal to $g_{m1,2}R_D$. At high frequencies, however, two effects degrade the balance of the phases: the parasitic capacitance at node P attenuates and delays the signal propagating from M_1 to M_2, and the gate-drain capacitance of M_1 provides a non-inverting feedforward path around M_1 (whereas M_2 does not contain such a path).

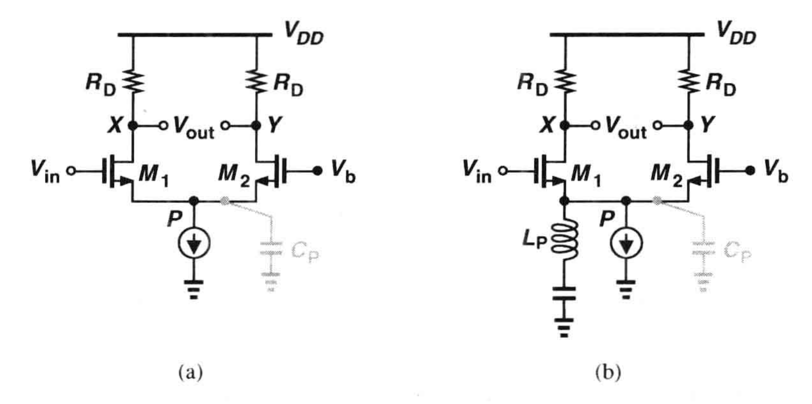


Figure 5.73 *Single-ended to differential conversion by (a) a simple differential pair, (b) a differential pair including tail resonance.*

The capacitance at P can be nulled through the use of a parallel inductor [Fig. 5.73(b)] [9], but the C_{GD1} feedforward persists. The tail inductor can be realized on-chip because its parallel equivalent resistance at resonance ($R_P = QL_P\omega_0$) is typically much greater than $1/g_{m1,2}$.

Example 5.26

A student computes C_P in Fig. 5.73(b) as $C_{SB1} + C_{SB2} + C_{GS2}$, and selects the value of L_P accordingly. Is this an appropriate choice?

Solution:

No, it is not. For L_P to null the phase shift at P, it must resonate with only $C_{SB1} + C_{SB2}$. This point can be seen by examining the voltage division at node P. As shown in Fig. 5.74, in the absence of $C_{SB1} + C_{SB2}$,

$$V_P = V_{in}\frac{Z_2}{Z_1 + Z_2}. \tag{5.158}$$

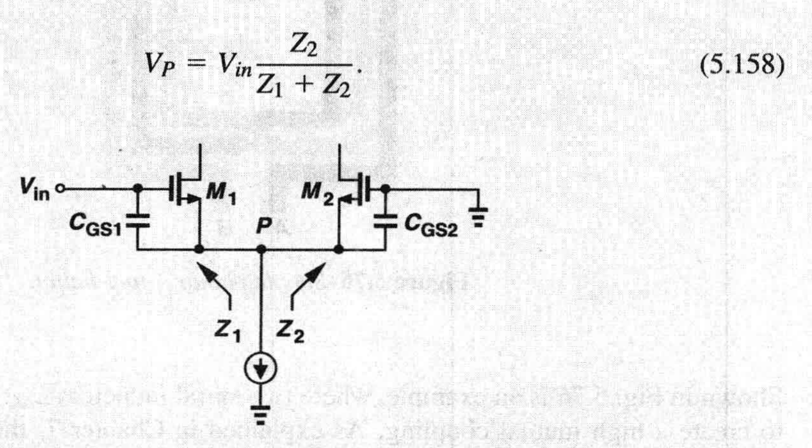


Figure 5.74 *Impedances seen at the common source of differential pair.*

For V_P to be exactly equal to half of V_{in} (with zero phase difference), we must have $Z_1 = Z_2$. Since each impedance is equal to $(g_m + g_{mb})^{-1}||(C_{GS}s)^{-1}$, we conclude that C_{GS2} must *not* be nulled.[20]

The topology of Fig. 5.73(b) still does not provide input matching. We must therefore insert (on-chip) inductances in series with the sources of M_1 and M_2 (Fig. 5.75). Here, L_{P1} and L_{P2} resonate with C_{P1} and C_{P2}, respectively, and $L_{S1} + L_{S2}$ provides the necessary input resistance. Of course, $L_{S1} + L_{S2}$ is realized as one inductor. However, as explained in Section 5.7, this topology exhibits a lower IP_3 than that of Fig. 5.65(b).

Balun Issues The foregoing development of differential LNAs has assumed ideal 1-to-1 baluns. Indeed, external baluns with a low loss (e.g., 0.5 dB) in the gigahertz range are available from manufacturers, but they consume board space and raise the cost. Integrated baluns, on the other hand, suffer from a relatively high loss and large capacitances.

20. But the parasitic capacitance of I_{SS} must be nulled.

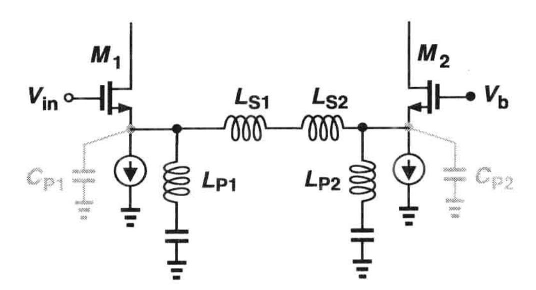


Figure 5.75 *Use of on-chip inductors for resonance and degeneration.*

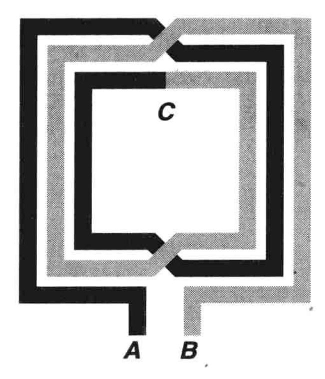


Figure 5.76 *Simple planar 1-to-1 balun.*

Shown in Fig. 5.76 is an example, where two spiral inductors L_{AC} and L_{CB} are intertwined to create a high mutual coupling. As explained in Chapter 7, the resistance and capacitance associated with the spirals and the sub-unity coupling factor make such baluns less attractive.

Example 5.27

A student attempts to use a 1-to-N balun with a differential CS stage so as to amplify the input voltage by a factor of N and potentially achieve a lower noise figure. Compute the noise figure in this case.

Solution:

Illustrated in Fig. 5.77, such an arrangement transforms the source impedance to a value of N^2R_S, requiring that each half circuit provide an input real part equal to $N^2R_S/2$. Thus, $L_1\omega_T = N^2R_S/2$, i.e., each degeneration inductance must be reduced by a factor of N^2. Since still half of the noise current of each input transistor flows to the output node, the noise power measured at each output is given by

$$\overline{V_{n,out1}^2} = \overline{V_{n,out2}^2} = 4kT\gamma g_{m1}\frac{R_1^2}{4} + 4kTR_1. \tag{5.159}$$

Example 5.27 *(Continued)*

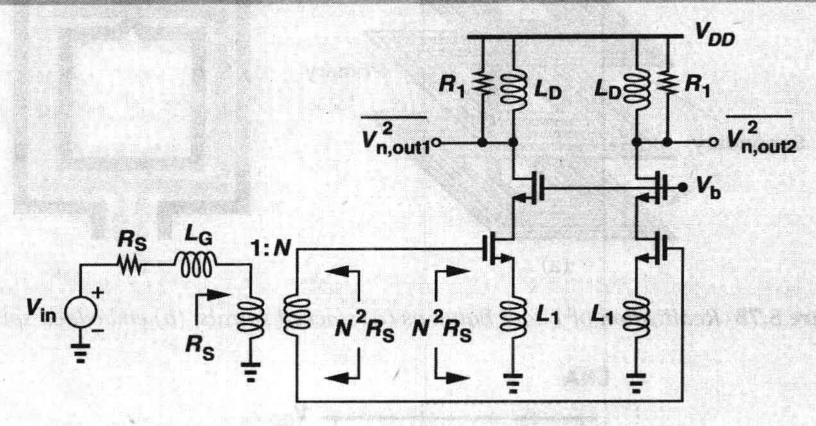


Figure 5.77 *Use of 1-to-N balun in an LNA.*

The gain from V_{in} to the differential output is now equal to $NR_1/(2L_1\omega_0)$. Doubling the above power, dividing by the square of the gain, and normalizing to $4kTR_S$, we have

$$\text{NF} = N^2\frac{\gamma}{2}g_{m1}R_S\left(\frac{\omega_0}{\omega_T}\right)^2 + 2N^2\frac{R_S}{R_1}\left(\frac{\omega_0}{\omega_T}\right)^2 + 1. \tag{5.160}$$

We note, with great distress, that the first two terms have *risen* by a factor of N^2![21] This is because the condition $L_1\omega_T = N^2R_S/2$ inevitably leads to an N^2-fold reduction in the transconductance of the circuit. Thus, even with the N-fold amplification of V_{in} by the balun, the overall voltage gain drops by a factor of N.

The reader may wonder if an N-to-1 (rather than 1-to-N) balun proves beneficial in the above example as it would multiply the first two terms of Eq. (5.160) by $1/N^2$ rather than N^2. Indeed, off-chip baluns may provide a lower noise figure if L_1 (a bond wire) can be reduced by a factor of N^2. On the other hand, on-chip baluns with a non-unity turns ratio are difficult to design and suffer from a higher loss and a lower coupling factor. Figure 5.78(a) shows an example [5], where one spiral forms the primary (secondary) of the balun and the series combination of two spirals constitutes the secondary (primary). Alternatively, as shown in Fig. 5.78(b), spirals having different numbers of turns can be embedded [10].

5.6.2 Other Methods of IP$_2$ Improvement

The difficulty with the use of off-chip or on-chip baluns at the input of differential LNAs makes single-ended topologies still an attractive choice. A possible approach to raising the IP$_2$ entails simply filtering the low-frequency second-order intermodulation product, called the beat component in Chapter 4. Illustrated in Fig. 5.79, the idea is to remove the beat by a simple high-pass filter (HPF) following the LNA. For example, suppose two interferers

21. Assuming that g_{m1} and ω_T remain unchanged.

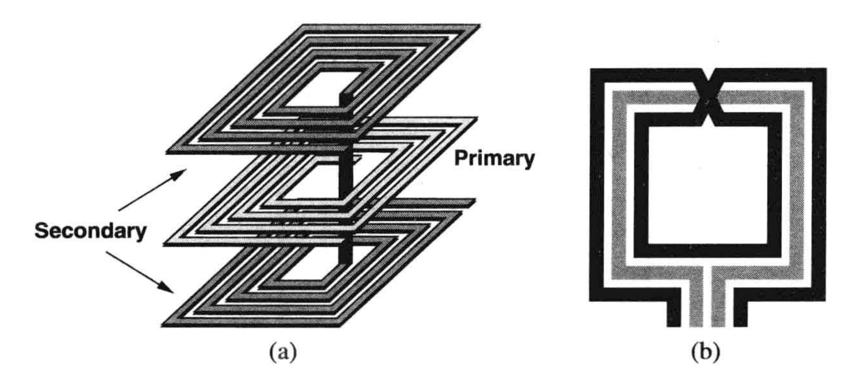


Figure 5.78 *Realization of 1-to-2 balun as (a) stacked spirals, (b) embedded spirals.*

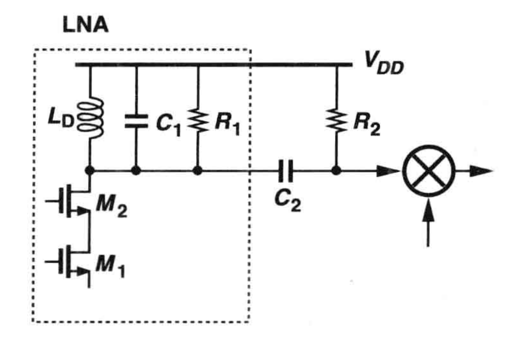


Figure 5.79 *Removal of low-frequency beat by first-order high-pass filter.*

are located at the edges of the 2.4-GHz band, $f_1 = 2.4$ GHz and $f_2 = 2.480$ GHz. The beat therefore lies at 80 MHz and is attenuated by approximately a factor of $2400/80 = 30$ for a first-order HPF. With this substantial suppression, the IP_2 of the LNA is unlikely to limit the RX performance, calling for techniques that improve the IP_2 of *mixers* (Chapter 6).

Example 5.28

A student considers the above calculation pessimistic, reasoning that an 80-MHz beat leaking to the baseband of an 11b/g or Bluetooth receiver does not fall within the desired channel. Is the student correct?

Solution:

Yes, the student is correct. For a direct-conversion 11b/g receiver, the baseband signal spans -10 MHz to $+10$ MHz. Thus, the worst-case beat occurs at 10 MHz, e.g., between two interferers at 2.400 GHz and 2.410 GHz. Such a beat is attenuated by a factor of $2400/10 = 240$ by the first-order HPF.

The filtration of the IM_2 product becomes less effective for wider communication bands. For example, if a receiver must accommodate frequencies from 1 GHz to 10 GHz, then two interferers can produce a beat *within* the band, prohibiting the use of filters to remove the beat. In this case, the LNA may become the receiver's IP_2 bottleneck.

5.7 NONLINEARITY CALCULATIONS

The general behavior of nonlinear systems was formulated in Chapter 2. In this section, we develop a methodology for computing the nonlinear characteristics of some circuits.

Recall from Chapter 2 that systems with weak static nonlinearity can be approximated by a polynomial such as $y = \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3$. Let us devise a method for computing α_1-α_3 for a given circuit. In many circuits, it is difficult to derive y as an explicit function of x. However, we recognize that

$$\alpha_1 = \frac{\partial y}{\partial x}|_{x=0} \tag{5.161}$$

$$\alpha_2 = \frac{1}{2}\frac{\partial^2 y}{\partial x^2}|_{x=0} \tag{5.162}$$

$$\alpha_3 = \frac{1}{6}\frac{\partial^3 y}{\partial x^3}|_{x=0}. \tag{5.163}$$

These expressions prove useful because we can obtain the derivatives by implicit differentiation. It is important to note that in most cases, $x = 0$ in fact corresponds to the *bias* point of the circuit with no input perturbation. In other words, the total y may not be zero for $x = 0$. For example, in the common-source stage of Fig. 5.80, M_1 is biased at a gate-source voltage of $V_{GS0} = V_b$ and V_{in} is superimposed on this voltage.

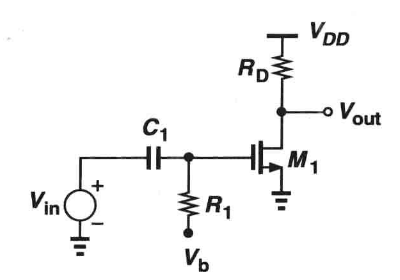


Figure 5.80 *CS stage with gate bias.*

5.7.1 Degenerated CS Stage

As an example, let us study the resistively-degenerated common-source stage shown in Fig. 5.81,

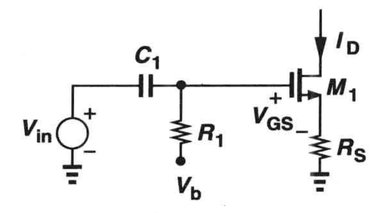


Figure 5.81 *CS stage for nonlinearity calculations.*

assuming the drain current is the output of interest. We wish to compute the IP_3 of the circuit. For a simple square-law device

$$I_D = K(V_{GS} - V_{TH})^2, \tag{5.164}$$

where $K = (1/2)\mu_n C_{ox}(W/L)$ and channel-length modulation and body effect are neglected. Since $V_{GS} = V_{in} - R_S I_D$,

$$I_D = K(V_{in} - R_S I_D - V_{TH})^2, \tag{5.165}$$

and hence

$$\frac{\partial I_D}{\partial V_{in}} = 2K(V_{in} - R_S I_D - V_{TH})\left(1 - R_S\frac{\partial I_D}{\partial V_{in}}\right). \tag{5.166}$$

We also note that

$$g_m = \frac{\partial I_D}{\partial V_{GS}} = 2K(V_{GS} - V_{TH}) \tag{5.167}$$

$$= 2K(V_{in0} - R_S I_{D0} - V_{TH}), \tag{5.168}$$

where V_{in0} ($= V_b$) and I_{D0} denote the bias values. Thus, in the absence of signals,

$$\frac{\partial I_D}{\partial V_{in}}|_{Vin0} = \alpha_1 = \frac{g_m}{1 + g_m R_S}, \tag{5.169}$$

an expected result.

We now compute the second derivative from Eq. (5.166):

$$\frac{\partial^2 I_D}{\partial V_{in}^2} = 2K\left(1 - R_S\frac{\partial I_D}{\partial V_{in}}\right)^2 + 2K(V_{in} - R_S I_D - V_{TH})\left(-R_S\frac{\partial^2 I_D}{\partial V_{in}^2}\right). \tag{5.170}$$

With no signals, (5.168) and (5.169) can be substituted in (5.170) to produce

$$\frac{\partial^2 I_D}{\partial V_{in}^2}|_{Vin0} = 2\alpha_2 = \frac{2K}{(1 + g_m R_S)^3}. \tag{5.171}$$

Lastly, we determine the third derivative from (5.170):

$$\frac{\partial^3 I_D}{\partial V_{in}^3} = 4K\left(1 - R_S\frac{\partial I_D}{\partial V_{in}}\right)\left(-R_S\frac{\partial^2 I_D}{\partial V_{in}^2}\right) + 2K\left(1 - R_S\frac{\partial I_D}{\partial V_{in}}\right)\left(-R_S\frac{\partial^2 I_D}{\partial V_{in}^2}\right)$$

$$- 2K(V_{in} - R_S I_D - V_{TH})R_S\frac{\partial^3 I_D}{\partial V_{in}^3}, \tag{5.172}$$

which, from (5.169) and (5.171) reduces to

$$\frac{\partial^3 I_D}{\partial V_{in}^3}|_{Vin0} = 6\alpha_3 = \frac{-12K^2 R_S}{(1 + g_m R_S)^5}. \tag{5.173}$$

While lengthy, the foregoing calculations lead to interesting results. Equation (5.173) reveals that $\alpha_3 = 0$ if $R_S = 0$, an expected outcome owing to the square-law behavior assumed for the transistor. Additionally, α_1 and α_3 have *opposite* signs, implying a compressive characteristic—whereas the undegenerated transistor would exhibit an *expansive* behavior. In other words, resistive degeneration of a square-law device *creates* third-order distortion.

To compute the IP_3 of the stage, we write from Chapter 2,

$$A_{IIP3} = \sqrt{\frac{4}{3}\left|\frac{\alpha_1}{\alpha_3}\right|} \tag{5.174}$$

$$= \sqrt{\frac{2g_m}{3R_S}\frac{(1+g_mR_S)^2}{K}}. \tag{5.175}$$

The 1-dB compression point follows the same expression but lowered by a factor of 3.03 (9.6 dB).

The reader may wonder if the above analysis of nonlinearity confuses large-signal and small-signal operations by expression α_1-α_3 in terms of the device transconductance. It is helpful to bear in mind that g_m in the above expressions is merely a short-hand notation for a *constant* value, $2K(V_{in0} - R_SI_{D0} - V_{TH})$, and independent of the input. It is, of course, plausible that α_1-α_3 must be independent of the input; otherwise, the polynomial's order exceeds 3.

Example 5.29

A student measures the IP_3 of the CS stage of Fig. 5.81 in the laboratory and obtains a value equal to *half* of that predicted by Eq. (5.175). Explain why.

Solution:

The test setup is shown in Fig. 5.82, where the signal generator produces the required input.[22] The discrepancy arises because the generator contains an internal output resistance $R_G = 50\,\Omega$, and it *assumes* that the circuit under test provides input matching, i.e., $Z_{in} = 50\,\Omega$. The generator's display therefore shows $A_0/2$ for the peak amplitude.

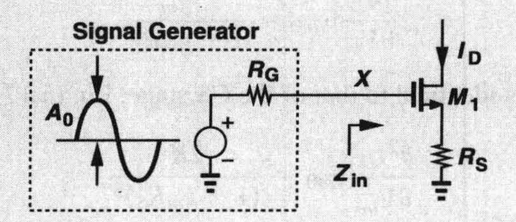


Figure 5.82 *CS stage driven by finite signal source impedance.*

(*Continues*)

22. In reality, the outputs of two generators are summed for a two-tone test.

Example 5.29 *(Continued)*

The simple CS stage, on the other hand, exhibits a high input impedance, sensing a peak amplitude of A_0 rather than $A_0/2$. Thus, the level that the student reads is half of that applied to the circuit. This confusion arises in IP_3 measurements because this quantity has been traditionally defined in terms of the *available* input power.

Example 5.30

Compute the IP_3 of a common-gate stage if the input is matched. Neglect channel-length modulation and body effect.

Solution:

The circuit is shown in Fig. 5.83, where we have

$$I_D = K(V_b - V_{in} - I_D R_S - V_{TH})^2, \tag{5.176}$$

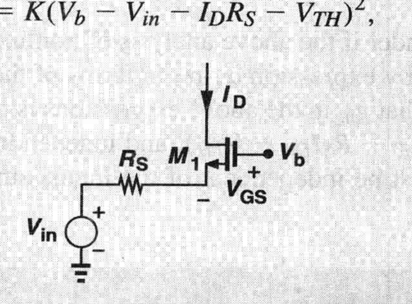


Figure 5.83 *CG stage for nonlinearity calculations.*

and $K = (1/2)\mu_n C_{ox}(W/L)$. Differentiating both sides with respect to V_{in} gives

$$\frac{\partial I_D}{\partial V_{in}} = 2K(V_b - V_{in} - I_D R_S - V_{TH})\left(-1 - R_S \frac{\partial I_D}{\partial V_{in}}\right). \tag{5.177}$$

In the absence of signals, $2K(V_b - V_{in0} - I_{D0}R_S - V_{TH})$ is equal to the transconductance of M_1, and hence

$$\frac{\partial I_D}{\partial V_{in}}|_{Vin0} = \frac{-g_m}{1 + g_m R_S}. \tag{5.178}$$

The second derivative is identical to that of the CS stage, Eq. (5.171):

$$\frac{\partial^2 I_D}{\partial V_{in}^2}|_{Vin0} = \frac{2K}{(1 + g_m R_S)^3}, \tag{5.179}$$

and the third derivative emerges as

$$\frac{\partial^3 I_D}{\partial V_{in}^3}|_{Vin0} = \frac{12K^2 R_S}{(1 + g_m R_S)^5}. \tag{5.180}$$

Example 5.30 *(Continued)*

Thus, the IP_3 expression in Eq. (5.175) applies here as well. For input matching, $R_S = 1/g_m$. However, as explained in Example 5.29, the definition of IP_3 is based on the *available* signal power, i.e., that which is delivered to a matched load. Thus, the peak value predicted by Eq. (5.175) must be divided by 2, yielding

$$A_{IIP3} = \frac{2}{K}\sqrt{\frac{2}{3}}g_m \quad (5.181)$$

$$= 4\sqrt{\frac{2}{3}}(V_{GS0} - V_{TH}), \quad (5.182)$$

where V_{GS0} denotes the bias value of the gate-source voltage.

5.7.2 Undegenerated CS Stage

Consider the CS stage shown in Fig. 5.80. Submicron transistors substantially depart from square-law characteristics. The effect of mobility degradation due to both vertical and lateral fields in the channel can be approximated as

$$I_D = \frac{1}{2}\mu_0 C_{ox}\frac{W}{L}\frac{(V_{GS} - V_{TH})^2}{1 + (\frac{\mu_0}{2v_{sat}L} + \theta)(V_{GS} - V_{TH})}, \quad (5.183)$$

where μ_0 denotes the zero-field mobility, v_{sat} the saturation velocity of the carriers, and θ the effect of the vertical field [11]. If the second term in the denominator remains much less than unity, we can write $(1 + \varepsilon)^{-1} \approx 1 - \varepsilon$ and hence

$$I_D \approx \frac{1}{2}\mu_0 C_{ox}\frac{W}{L}\left[(V_{GS} - V_{TH})^2 - \left(\frac{\mu_0}{2v_{sat}L} + \theta\right)(V_{GS} - V_{TH})^3\right]. \quad (5.184)$$

The input signal, V_{in}, is superimposed on a bias voltage, $V_{GS0} = V_b$. We therefore replace V_{GS} with $V_{in} + V_{GS0}$, obtaining

$$I_D \approx K[2 - 3a(V_{GS0} - V_{TH})](V_{GS0} - V_{TH})V_{in} + K[1 - 3a(V_{GS0} - V_{TH})]V_{in}^2$$
$$- KaV_{in}^3 + K(V_{GS0} - V_{TH})^2 - aK(V_{GS0} - V_{TH})^3, \quad (5.185)$$

where $K = (1/2)\mu_0 C_{ox}(W/L)$ and $a = \mu_0/(2v_{sat}L) + \theta$. We recognize the coefficient of V_{in} as the transconductance $(\partial I_D/\partial V_{in})$ and the last two terms as the bias current. It follows that

$$\alpha_1 = K[2 - 3a(V_{GS0} - V_{TH})](V_{GS0} - V_{TH}) \quad (5.186)$$

$$\alpha_3 = -Ka. \quad (5.187)$$

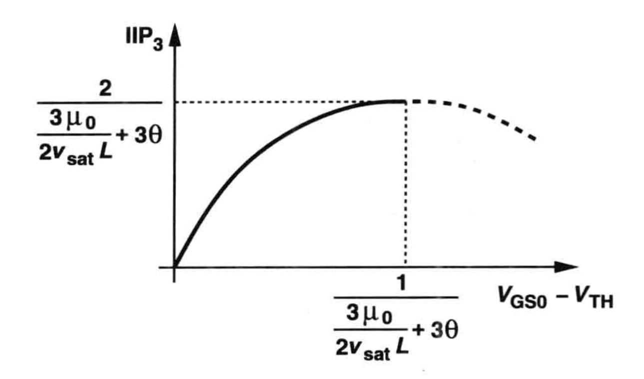


Figure 5.84 *Behavior of IP_3 as a function of overdrive.*

The IP_3 is given by

$$A_{IIP3} = \sqrt{\frac{4}{3} \times \frac{2 - 3a(V_{GS0} - V_{TH})}{a}(V_{GS0} - V_{TH})} \tag{5.188}$$

$$= \sqrt{\frac{\frac{8}{3}(V_{GS0} - V_{TH})}{\frac{\mu_0}{2v_{sat}L} + \theta} - 4(V_{GS0} - V_{TH})^2}. \tag{5.189}$$

We note that the IP_3 rises with the bias overdrive voltage, reaching a maximum of

$$A_{IIP3,max} = \frac{2}{3a} = \frac{2}{3}\frac{1}{\frac{\mu_0}{2v_{sat}L} + \theta} \tag{5.190}$$

at $V_{GS0} - V_{TH} = (3a)^{-1}$ (Fig. 5.84).

Example 5.31

If the second term in the denominator of Eq. (5.183) is only somewhat less than unity, a better approximation must be used, e.g., $(1 + \varepsilon)^{-1} \approx 1 - \varepsilon + \varepsilon^2$. Compute α_1 and α_3 with this approximation.

Solution:

The additional term $a^2(V_{GS} - V_{TH})^2$ is multiplied by $K(V_{GS} - V_{TH})^2$, yielding two terms of interest: $4Ka^2V_{in}(V_{GS} - V_{TH})^3$ and $4Ka^2V_{in}^3(V_{GS} - V_{TH})$. The former contributes to α_1 and the latter to α_3. It follows that

$$\alpha_1 = K[2 - 3a(V_{GS0} - V_{TH}) + 4a^2(V_{GS0} - V_{TH})^2](V_{GS0} - V_{TH}) \tag{5.191}$$

$$\alpha_3 = -Ka[1 - 4a(V_{GS0} - V_{TH})]. \tag{5.192}$$

5.7.3 Differential and Quasi-Differential Pairs

In RF systems, differential signals can be processed using the differential pair shown in Fig. 5.85(a) or the "quasi-differential" pair depicted in Fig. 5.85(b). The two topologies exhibit distinctly different nonlinear characteristics. We know from our above analysis that the dependence of the mobility upon vertical and lateral fields in the channel results in third-order nonlinearity in the quasi-differential pair and an IP_3 given by Eq. (5.189). To study the nonlinearity of the standard differential pair, we recall from basic analog circuits that

$$I_{D1} - I_{D2} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}V_{in}\sqrt{\frac{4I_{SS}}{\mu_n C_{ox}\frac{W}{L}} - V_{in}^2}, \tag{5.193}$$

where V_{in} denotes the input differential voltage. If $|V_{in}| \ll I_{SS}/(\mu_n C_{ox}W/L)$, then

$$I_{D1} - I_{D2} \approx \frac{1}{2}\mu_n C_{ox}\frac{W}{L}V_{in}\sqrt{\frac{4I_{SS}}{\mu_n C_{ox}\frac{W}{L}}}\left(1 - \frac{1}{2}\frac{V_{in}^2}{\frac{4I_{SS}}{\mu_n C_{ox}W/L}}\right). \tag{5.194}$$

That is,

$$\alpha_1 = \sqrt{\mu_n C_{ox}\frac{W}{L}I_{SS}} \tag{5.195}$$

$$\alpha_3 = -\left(\mu_n C_{ox}\frac{W}{L}\right)^{3/2}\frac{1}{8\sqrt{I_{SS}}}, \tag{5.196}$$

and hence

$$A_{IIP3} = \sqrt{\frac{6I_{SS}}{\mu_n C_{ox}W/L}} \tag{5.197}$$

$$= \sqrt{6}(V_{GS0} - V_{TH}), \tag{5.198}$$

where $(V_{GS0} - V_{TH})$ is the overdrive voltage of each transistor in equilibrium ($V_{in} = 0$).[23]

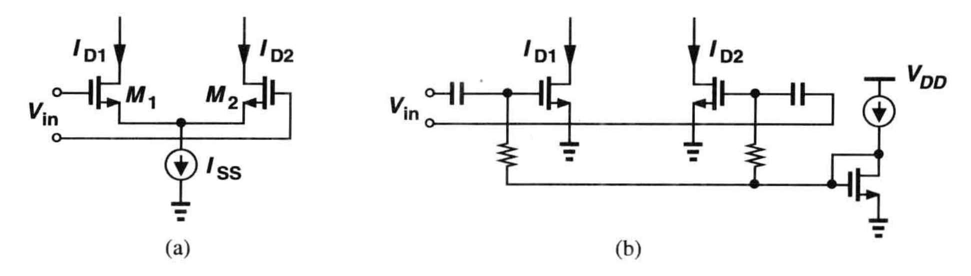


Figure 5.85 *(a) Differential and (b) quasi-differential pairs.*

23. Note that one transistor turns off if the differential input reaches $\sqrt{2}(V_{GS0} - V_{TH})$.

Interestingly, the standard differential pair suffers from third-order nonlinearity even in the absence of field-dependent mobility (i.e., with square-law devices). For this reason, the quasi-differential pair of Fig. 5.85(b) is preferred in cases where linearity is important. In fact, it is for this reason that the differential CS LNA of Fig. 5.65(b) does not employ a tail current source. The quasi-differential pair also saves the voltage headroom associated with the tail current source, proving more attractive as the supply voltage is scaled down.

5.7.4 Degenerated Differential Pair

Consider the degenerated pair shown in Fig. 5.86, where $I_{D1} - I_{D2}$ is the output of interest. Since $I_{D1} + I_{D2} = 2I_0$, we have $\partial(I_{D1} - I_{D2})/\partial V_{in} = 2\partial I_{D1}/\partial V_{in}$. Also, $V_{in1} - V_{GS1} - I_S R_S = V_{in2} - V_{GS2}$ and $I_S = I_{D1} - I_0$. It follows that

$$V_{in} - R_S I_{D1} + R_S I_0 = \frac{1}{\sqrt{K}}(\sqrt{I_{D1}} - \sqrt{I_{D2}}), \tag{5.199}$$

where $V_{in} = V_{in1} - V_{in2}$ and $K = (1/2)\mu_n C_{ox}(W/L)$. Differentiating both sides with respect to V_{in} yields

$$\frac{\partial I_{D1}}{\partial V_{in}}\left[R_S + \frac{1}{2\sqrt{K}}\left(\frac{1}{\sqrt{I_{D1}}} + \frac{1}{\sqrt{I_{D2}}}\right)\right] = 1. \tag{5.200}$$

At $V_{in} = 0$, $I_{D1} = I_{D2}$ and

$$\alpha_1 = \frac{1}{R_S + \dfrac{2}{g_m}}, \tag{5.201}$$

where $g_m = 2I_0/(V_{GS0} - V_{TH})$. Differentiating both sides of (5.200) with respect to V_{in} gives

$$\frac{\partial^2 I_{D1}}{\partial V_{in}^2}\left[R_S + \frac{1}{2\sqrt{K}}\left(\frac{1}{\sqrt{I_{D1}}} + \frac{1}{\sqrt{I_{D2}}}\right)\right] - \frac{\partial I_{D1}}{\partial V_{in}}\left[\frac{1}{4\sqrt{K}}\left(\frac{1}{I_{D1}^{3/2}}\frac{\partial I_{D1}}{\partial V_{in}} + \frac{1}{I_{D2}^{3/2}}\frac{\partial I_{D2}}{\partial V_{in}}\right)\right] = 0. \tag{5.202}$$

Note that for $V_{in} = 0$, we have $\partial^2 I_{D1}/\partial V_{in}^2 = 0$ because $\partial I_{D1}/\partial V_{in} = -\partial I_{D2}/\partial V_{in}$ and the term in the second set of square brackets vanishes. Differentiating once more and exploiting

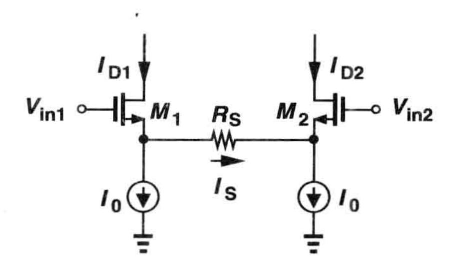


Figure 5.86 *Degenerated differential pair.*

this fact, we have

$$\frac{\partial^3 I_{D1}}{\partial V_{in}^3}|_{Vin=0} = \frac{-3}{(R_S + \frac{2}{g_m})^4 g_m I_0^2} = 6\alpha_3. \tag{5.203}$$

It follows that $6\alpha_3 = \partial^3(I_{D1} - I_{D2})/\partial V_{in}^3 = 2\partial^3 I_{D1}/\partial V_{in}^3$. We now have that

$$A_{IIP3} = \frac{2I_0}{3}\sqrt{g_m\left(R_S + \frac{2}{g_m}\right)^3}. \tag{5.204}$$

REFERENCES

[1] J. Rogin et al., "A 1.5-V 45-mW Direct-Conversion WCDMA Receiver IC in 0.13-m CMOS," *IEEE Journal of Solid-State Circuits,* vol. 38, pp. 2239–2248, Dec. 2003.

[2] D. K. Shaeffer and T. H. Lee, "A 1.5-V, 1.5-GHz CMOS Low Noise Amplifier," *IEEE J. Solid-State Circuits,* vol. 32, pp. 745–759, May 1997.

[3] P. Rossi et al., "A Variable-Gain RF Front End Based on a Voltage-Voltage Feedback LNA for Multistandard Applications," *IEEE J. Solid-State Circuits,* vol. 40, pp. 690–697, March 2005.

[4] X. Li, S. Shekar, and D. J. Allstot, "G_m-Boosted Common-Gate LNA and Differential Colpitts VCO/QVCO in 0.18-um CMOS," *IEEE J. Solid-State Circuits,* vol. 40, pp. 2609–2618, Dec. 2005.

[5] A. Zolfaghari, A. Y. Chan, and B. Razavi, "Stacked Inductors and 1-to-2 Transformers in CMOS Technology," *IEEE Journal of Solid-State Circuits,* vol. 36, pp. 620–628, April 2001.

[6] F. Bruccoleri, E. A. M. Klumperink, and B. Nauta, "Wideband CMOS Low-Noise Amplifier Exploiting Thermal Noise Canceling," *IEEE J. Solid-State Circuits,* vol. 39, pp. 275–281, Feb. 2004.

[7] B. Razavi, "Cognitive Radio Design Challenges and Techniques," *IEEE Journal of Solid-State Circuits,* vol. 45, pp. 1542–1553, Aug. 2010.

[8] B. Razavi et al., "A UWB CMOS Transceiver," *IEEE Journal of Solid-State Circuits,* vol. 40, pp. 2555–2562, Dec. 2005.

[9] M. Zargari et al., "A Single-Chip Dual-Band Tri-Mode CMOS Transceiver for IEEE 802.11a/b/g Wireless LAN," *IEEE Journal of Solid-State Circuits,* vol. 39, pp. 2239–2249, Dec. 2004.

[10] J. R. Long and M. A. Copeland, "The Modeling, Characterization, and Design of Monolithic Inductors for Silicon RF ICs," *IEEE J. Solid-State Circuits,* vol. 32, pp. 357–369, March 1997.

[11] B. Razavi, *Design of Analog CMOS Integrated Circuits,* Boston: McGraw-Hill, 2001.

PROBLEMS

5.1. Assuming $Z_{in} = x + jy$, derive an equation for constant-Γ contours in Fig. 5.4.

5.2. If $R_p = R_S$ and $g_m R_S \approx 1$, determine the NF in Eq. (5.18) by considering the first three terms. What value of g_m is necessary to achieve a noise figure of 3.5 dB?

5.3. Repeat Example 5.5 by solving the specific network shown in Fig. 5.10(a).

5.4. Determine the noise figure of the stages shown in Fig. 5.87 with respect to a source impedance of R_S. Neglect channel-length modulation and body effect.

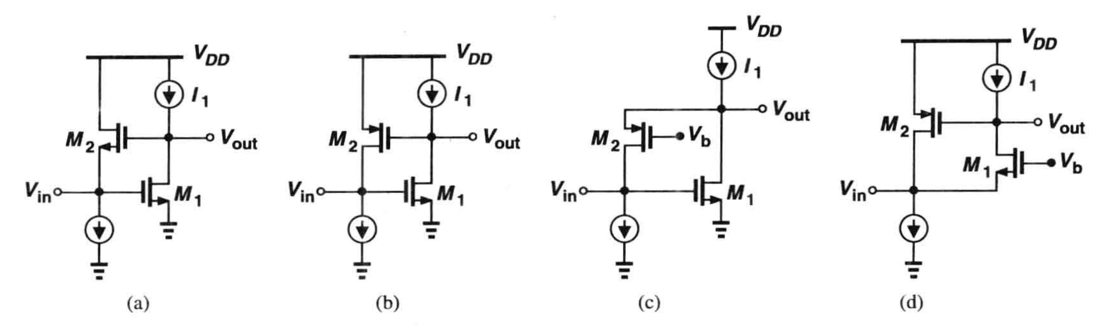


Figure 5.87 *Stages for NF calculation.*

5.5. For the inductively-loaded CS stage of Fig. 5.11(b), determine V_{out}/V_{in} and find the voltage gain at the resonance frequency, $\omega_0 = 1/\sqrt{L_1(C_1 + C_F)}$, if $|jC_1\omega_0| \ll g_m$.

5.6. For the CS stage of Fig. 5.13(a), determine the closed-loop gain and noise figure if channel-length modulation is not neglected. Assume matching at the input.

5.7. For the complementary stage shown in Fig. 5.15, determine the closed-loop gain and noise figure if channel-length modulation is not neglected. Assume matching at the input.

5.8. For the CG stage of Fig. 5.16(a), compute the noise figure at the output resonance frequency if $g_m \neq 1/R_S$. How can g_m be chosen to yield a noise figure lower than $1 + \gamma + 4R_S/R_a$?

5.9. A circuit exhibits a noise figure of 3 dB. What percentage of the output noise power is due to the source resistance, R_S? Repeat the problem for NF = 1 dB.

5.10. Determine the noise figure of the CG circuits shown in Fig. 5.17.

5.11. In Example, 5.10, we concluded that the noise of M_2 reaches the output unattenuated if ω is greater than $(R_1C_X)^{-1}$ but much less than $g_{m2}/(C_{GS2} + C_X)$. Does such a frequency range exist? In other words, under what conditions do we have $(R_1C_X)^{-1} < \omega \ll g_{m2}/(C_{GS2} + C_X) >$? Assume $g_{m2} \approx g_{m1}$ and recall that $g_{m1}R_1$ is the gain of the LNA and C_X is on the order of C_{SG2}.

5.12. If L_G in Fig. 5.34 suffers from a series resistance of R_t, determine the noise figure of the circuit.

5.13. The LNA shown in Fig. 5.88 is designed to operate with low supply voltages. Each inductor is chosen to resonate with the total capacitance at its corresponding node at the frequency of interest. Neglect channel-length modulation and body effect and the noise due to the loss in L_2. Determine the noise figure of the LNA with respect to a source resistance R_S assuming that L_1 can be viewed as a resistance equal to R_p at the resonance frequency. Make sure the result reduces to a familiar form if $R_p \to \infty$. (Hint: the equivalent transconductance of a degenerated common-source stage is given by $g_m/(1 + g_mR_1)$, where R_1 denotes the degeneration resistance.)

5.14. Determine S_{11} for both topologies in Fig. 5.40 and compute the maximum deviation of the center frequency for which S_{11} remains lower than -10 dB.

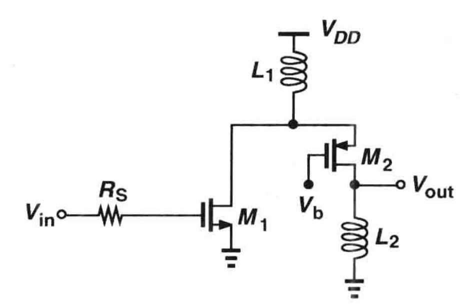


Figure 5.88 *Folded-cascode LNA.*

5.15. Repeat the analysis of the CG stage in Fig. 5.43 while including channel-length modulation.

5.16. Repeat the NF analysis of the CG stage in Fig. 5.43 while including the noise of the feedback network as a voltage, $\overline{V_{nF}^2}$, in series with its input.

5.17. Prove that the input-referred noise of the feedforward amplifier in Fig. 5.45(a) manifests itself as the fourth term in Eq. (5.124).

5.18. Repeat the analysis of the CG stage of Fig. 5.45(a) while including channel-length modulation.

5.19. For the circuit shown in Fig. 5.89, we express the input-output characteristic as

$$I_{out} - I_0 = \alpha_1(V_{in} - V_0) + \alpha_2(V_{in} - V_0)^2 + \cdots, \quad (5.205)$$

where I_0 and V_0 denote the bias values, i.e., the values in the absence of signals. We note that $\partial I_{out}/\partial V_{in} = \alpha_1$ at $V_{in} = V_0$ (or $I_{out} = I_0$). Similarly, $\partial I_{out}^2/\partial^2 V_{in} = 2\alpha_2$ at $V_{in} = V_0$ (or $I_{out} = I_0$).

(a) Write a KVL around the input network in terms of V_{in} and I_{out} (with no V_{GS}). Differentiate both sides *implicitly* with respect to V_{in}. You will need this equation in part (b). Noting that $2\sqrt{KI_0} = g_m$, where $K = \mu_n C_{ox} W/L$, find $\partial I_{out}/\partial V_{in}$ and hence α_1.

(b) Differentiate the equation obtained in part (a) with respect to V_{in} once more and compute α_2 in terms of I_0 and g_m.

(c) Determine the IP_2 of the circuit.

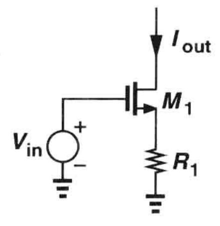


Figure 5.89 *Stage for IP_2 calculation.*

5.20. Determine the noise figure in Example 5.21 if the gain is reduced by 3 dB.

5.21. Compare the power consumptions of the single-ended and differential CS stages discussed in Section 5.6.1. Consider two cases: (a) the differential stage is derived by only halving L_1 (and hence has a lower noise figure), or (b) the differential stage is designed for the same NF as the single-ended circuit.

5.22. Repeat the analysis of the differential CG stage NF if a 1-to-2 balun is used. Such a balun provides a voltage gain of 2.

5.23. Consider a MOS transistor configured as a CS stage and operating in saturation. Determine the IP_3 and P_{1dB} if the device (a) follows the square-law behavior, $I_D \propto (V_{GS} - V_{TH})^2$, or (b) exhibits field-dependent mobility [Eq. (5.183)]. (Hint: IP_3 and P_{1dB} may not be related by a 9.6-dB difference in this case.)

CHAPTER



MIXERS

In this chapter, our study of building blocks focuses on downconversion and upconversion mixers, which appear in the receive path and the transmit path, respectively. While a decade ago, most mixers were realized as a Gilbert cell, many more variants have recently been introduced to satisfy the specific demands of different RX or TX architectures. In other words, a stand-alone mixer design is no longer meaningful because its ultimate performance heavily depends on the circuits surrounding it. The outline of the chapter is shown below.

General Considerations	Passive Mixers	Active Mixers	Improved Mixer Topologies	Upconversion Mixers
Mixer Noise Figures	Conversion Gain	Conversion Gain	Active Mixers with Current Source Helpers	Passive Mixers
Port-to-Port Feedthrough	Noise	Noise	Active Mixers with High IP_2	Active Mixers
Single-Balanced and Double-Balanced Mixers	Input Impedance	Linearity	Active Mixers with Low Flicker Noise	
Passive and Active Mixers	Current-Driven Mixers			

6.1 GENERAL CONSIDERATIONS

Mixers perform frequency translation by multiplying two waveforms (and possibly their harmonics). As such, mixers have three distinctly different ports. Figure 6.1 shows a generic transceiver environment in which mixers are used. In the receive path, the downconversion mixer senses the RF signal at its "RF port" and the local oscillator waveform at its "LO port." The output is called the "IF port" in a heterodyne RX or the "baseband port" in a direct-conversion RX. Similarly, in the transmit path, the upconversion mixer input sensing the IF or the baseband signal is called the IF port or the baseband port, and the output port is called the RF port. The input driven by the LO is called the LO port.

How linear should each input port of a mixer be? A mixer can simply be realized as depicted in Fig. 6.2(a), where V_{LO} turns the switch on and off, yielding $V_{IF} = V_{RF}$ or

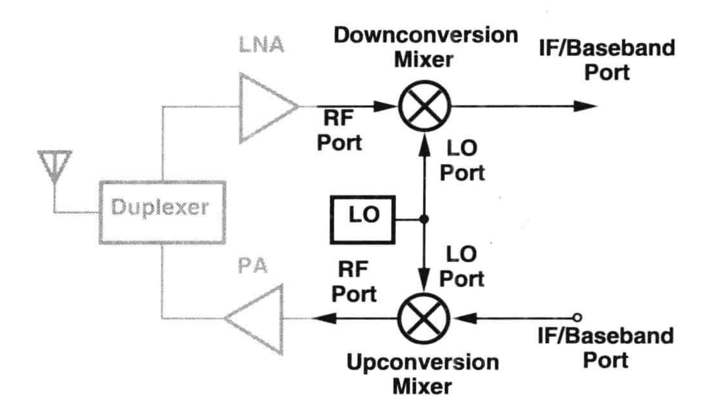


Figure 6.1 *Role of mixers in a generic transceiver.*

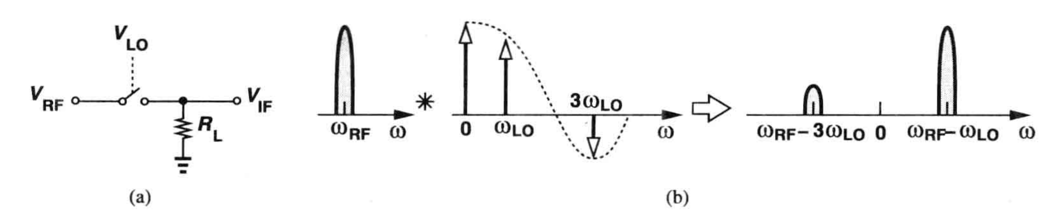


Figure 6.2 *(a) Mixer using an ideal switch, (b) input and output spectra.*

$V_{IF} = 0$. As explained in Chapter 2, with abrupt switching, the operation can be viewed as multiplication of the RF input by a square wave toggling between 0 and 1, even if V_{LO} itself is a *sinusoid*. Thus, as illustrated in Fig. 6.2(b), the circuit mixes the RF input with all of the LO harmonics, producing what we called "mixing spurs" in Chapter 4. In other words, the LO port of this mixer is very nonlinear. The RF port, of course, must remain sufficiently linear to satisfy the compression and/or intermodulation requirements.

The reader may wonder if the LO port of mixers can be linearized so as to avoid mixing with the LO harmonics. As seen later in this chapter, mixers suffer from a lower gain and higher noise as the switching in the LO port becomes less abrupt. We therefore design mixers and LO swings to ensure abrupt switching and deal with mixing spurs at the architecture level (Chapter 4).

6.1.1 Performance Parameters

Let us now consider mixer performance parameters and their role in a transceiver.

Noise and Linearity In a receive chain, the input noise of the mixer following the LNA is divided by the LNA gain when referred to the RX input. Similarly, the IP_3 of the mixer is scaled down by the LNA gain. (Recall from Chapter 5 that the mixer noise and IP_3 are divided by *different* gains.) The design of downconversion mixers therefore entails a compromise between the noise figure and the IP_3 (or P_{1dB}). Also, the designs of the LNA and the mixer are inextricably linked, requiring that the cascade be designed as one entity.

Where in the design space do we begin then? Since the noise figure of mixers is rarely less than 8 dB, we typically allocate a gain of 10 to 15 dB to the LNA and proceed with

the design of the mixer, seeking to maximize its linearity while not raising its NF. If the resulting mixer design is not satisfactory, some iteration becomes necessary. For example, we may decide to further linearize the mixer even if the NF increases and compensate for the higher noise by raising the LNA gain. We elaborate on these points in various design examples in this chapter.

In direct-conversion receivers, the IP_2 of the LNA/mixer cascade must be maximized. In Section 6.4, we introduce methods of raising the IP_2 in mixers. Also, as mentioned in Chapter 4, the mixing spurs due to the LO harmonics become important in broadband receivers.

For upconversion mixers, the noise proves somewhat critical only if the TX output noise in *the RX band* must be very small (Chapter 4), but even such cases demand more relaxed mixer noise performance than receivers do. The linearity of upconversion mixers is specified by the type of modulation and the baseband signal swings.

Gain Downconversion mixers must provide sufficient gain to adequately suppress the noise contributed by subsequent stages. However, low supply voltages make it difficult to achieve a gain of more than roughly 10 dB while retaining linearity. Thus, the noise of stages following the mixer still proves critical.

In direct-conversion transmitters, it is desirable to maximize the gain and hence the output swings of upconversion mixers, thereby relaxing the gain required of the power amplifier. In two-step transmitters, on the other hand, the IF mixers must provide only a moderate gain so as to avoid compressing the RF mixer.

The gain of mixers must be carefully defined to avoid confusion. The "voltage conversion gain" of a downconversion mixer is given by the ratio of the rms voltage of the IF signal to the rms voltage of the RF signal. Note that these two signals are centered around two different frequencies. The voltage conversion gain can be measured by applying a sinusoid at ω_{RF} and finding the amplitude of the downconverted component at ω_{IF}. For upconversion mixers, the voltage conversion gain is defined in a similar fashion but from the baseband or IF port to the RF port.

In traditional RF and microwave design, mixers are characterized by a "power conversion gain," defined as the output signal power divided by the input signal power. But in modern RF design, we prefer to employ voltage quantities because the input impedances are mostly imaginary, making the use of power quantities difficult and unnecessary.

Port-to-Port Feedthrough Owing to device capacitances, mixers suffer from unwanted coupling (feedthrough) from one port to another [Fig. 6.3(a)]. For example, if the mixer

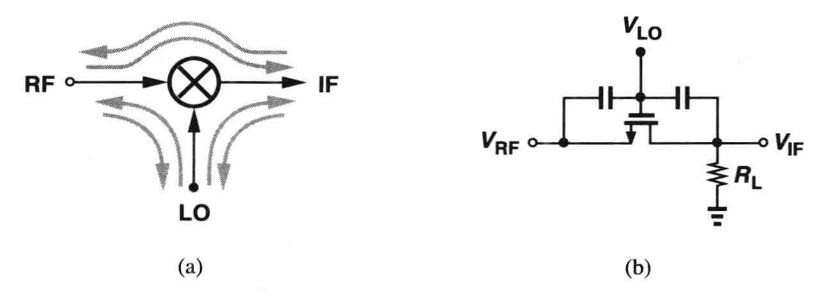


Figure 6.3 *(a) Feedthrough mechanisms in a mixer, (b) feedthrough paths in a MOS mixer.*

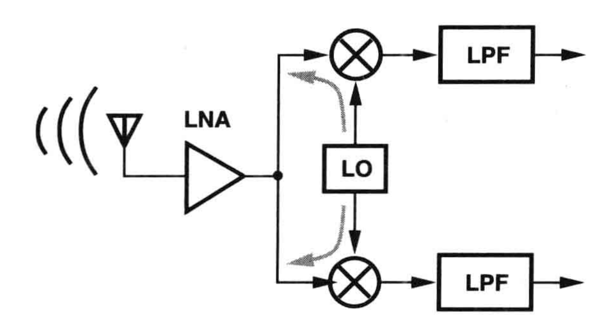


Figure 6.4 *Effect of LO-RF feedthrough.*

is realized by a MOSFET [Fig. 6.3(b)], then the gate-source and gate-drain capacitances create feedthrough from the LO port to the RF and IF ports.

The effect of mixer port-to-port feedthrough on the performance depends on the architecture. Consider the direct-conversion receiver shown in Fig. 6.4. As explained in Chapter 4, the LO-RF feedthrough proves undesirable as it produces both offsets in the baseband and LO radiation from the antenna. Interestingly, this feedthrough is entirely determined by the symmetry of the mixer circuit and LO waveforms (Section 6.2.2). The LO-IF feedthrough is benign because it is heavily suppressed by the baseband low-pass filter(s).

Example 6.1

Consider the mixer shown in Fig. 6.5, where $V_{LO} = V_1 \cos \omega_{LO} t + V_0$ and C_{GS} denotes the gate-source overlap capacitance of M_1. Neglecting the on-resistance of M_1 and assuming abrupt switching, determine the dc offset at the output for $R_S = 0$ and $R_S > 0$. Assume $R_L \gg R_S$.

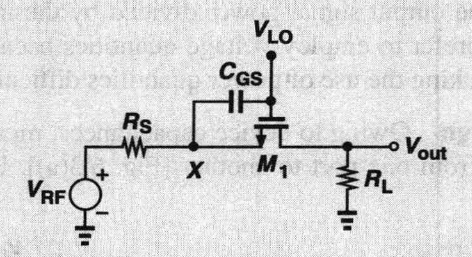


Figure 6.5 *LO-RF feedthrough in a MOS device operating as a mixer.*

Solution:

The LO leakage to node X is expressed as

$$V_X = \frac{R_S C_{GS} s}{R_S C_{GS} s + 1} V_{LO}, \tag{6.1}$$

Example 6.1 (*Continued*)

because even when M_1 is on, node X sees a resistance of approximately R_S to ground. With abrupt switching, this voltage is multiplied by a square wave toggling between 0 and 1. The output dc offset results from the mixing of V_X and the first harmonic of the square wave. Exhibiting a magnitude of $2\sin(\pi/2)/\pi = 2/\pi$, this harmonic can be expressed as $(2/\pi)\cos\omega_{LO}t$, yielding

$$V_{out}(t) = V_X(t) \times \frac{2}{\pi}\cos\omega_{LO}t + \cdots \tag{6.2}$$

$$= \frac{R_S C_{GS}\omega_{LO}}{\sqrt{R_S^2 C_{GS}^2 \omega_{LO}^2 + 1}} V_1 \cos(\omega_{LO}t + \phi) \times \frac{2}{\pi}\cos\omega_{LO}t + \cdots, \tag{6.3}$$

where $\phi = (\pi/2) - \tan^{-1}(R_S C_{GS}\omega_{LO})$. The dc component is therefore equal to

$$V_{dc} = \frac{V_1}{\pi}\frac{R_S C_{GS}\omega_{LO}\cos\phi}{\sqrt{R_S^2 C_{GS}^2 \omega_{LO}^2 + 1}}. \tag{6.4}$$

As expected, the output dc offset vanishes if $R_S = 0$.

The generation of dc offsets can also be seen intuitively. Suppose, as shown in Fig. 6.6, the RF input is a sinusoid having the same frequency as the LO. Then, each time the switch turns on, the *same* portion of the input waveform appears at the output, producing a certain average.

The RF-LO and RF-IF feedthroughs also prove problematic in direct-conversion receivers. As shown in Fig. 6.7, a large in-band interferer can couple to the LO and injection-pull it (Chapter 8), thereby corrupting the LO spectrum. To avoid this effect,

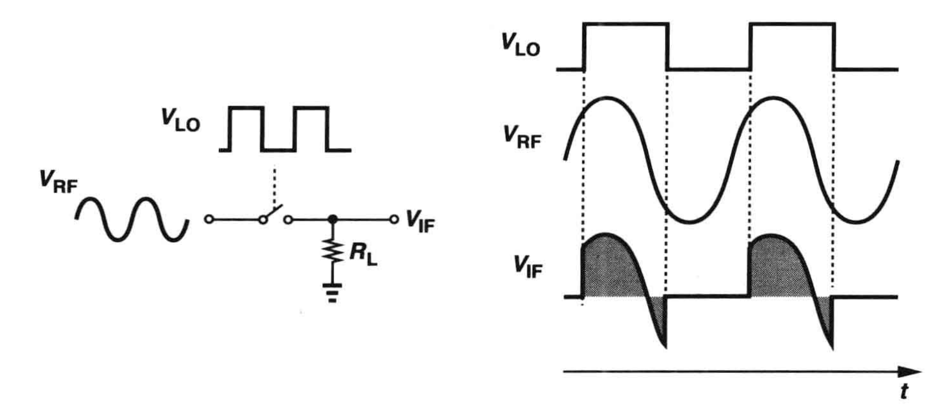


Figure 6.6 *Offset generated by LO leakage.*

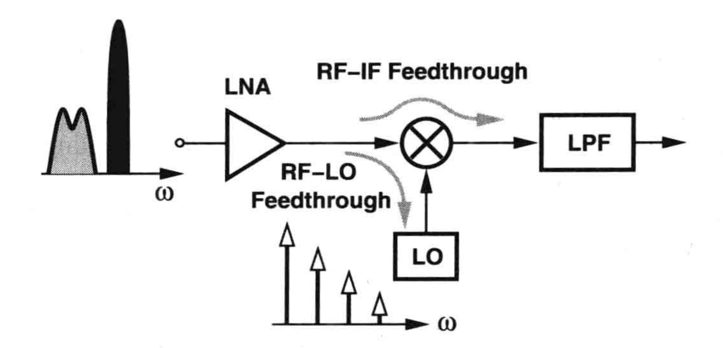


Figure 6.7 *Effect of RF-LO feedthrough in a direct-conversion receiver.*

a buffer is typically interposed between the LO and the mixer. Also, as explained in Chapter 4, the RF-IF feedthrough corrupts the baseband signal by the beat component resulting from even-order distortion in the RF path. (This phenomenon is characterized by the IP_2.)

Now, consider the heterodyne RX depicted in Fig. 6.8. Here, the LO-RF feedthrough is relatively unimportant because (1) the LO leakage falls outside the band and is attenuated by the selectivity of the LNA, the front-end band-select filter, and the antenna; and (2) the dc offset appearing at the output of the RF mixer can be removed by a high-pass filter. The LO-IF feedthrough, on the other hand, becomes serious if ω_{IF} and ω_{LO} are too close to allow filtering of the latter. The LO feedthrough may then desensitize the IF mixers if its level is comparable with their 1-dB compression point.

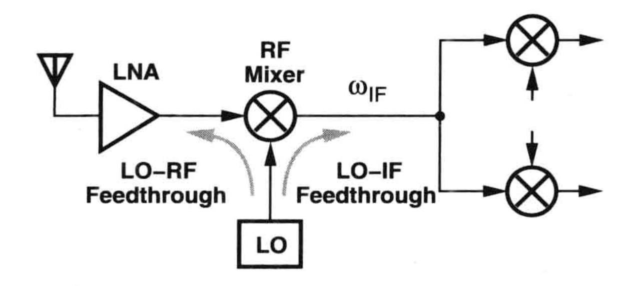


Figure 6.8 *Effect of LO feedthrough in a heterodyne RX.*

Example 6.2

Shown in Fig. 6.9 is a receiver architecture wherein $\omega_{LO} = \omega_{RF}/2$ so that the RF channel is translated to an IF of $\omega_{RF} - \omega_{LO} = \omega_{LO}$ and subsequently to zero. Study the effect of port-to-port feedthroughs in this architecture.

Example 6.2 (*Continued*)

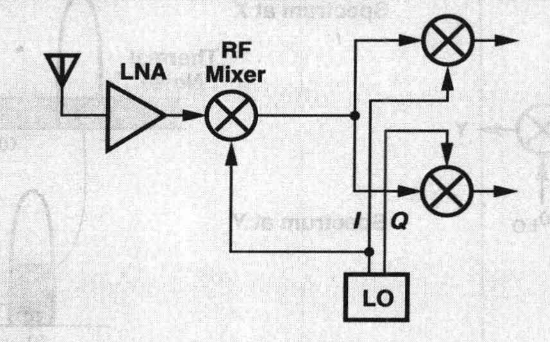


Figure 6.9 *Half-RF RX architecture.*

Solution:

For the RF mixer, the LO-RF feedthrough is unimportant as it lies at $\omega_{RF}/2$ and is suppressed. Also, the RF-LO feedthrough is not critical because in-band interferers are far from the LO frequency, creating little injection pulling. (Interferers near the LO frequency are attenuated by the front end before reaching the mixer.) The RF-IF feedthrough proves benign because low-frequency beat components appearing at the RF port can be removed by high-pass filtering.

The most critical feedthrough in this architecture is that from the LO port to the IF port of the RF mixer. Since $\omega_{IF} = \omega_{LO}$, this leakage lies in the *center* of the IF channel, potentially desensitizing the IF mixers (and producing dc offsets in the baseband). Thus, the RF mixer must be designed for minimal LO-IF feedthrough (Section 6.1.3).

The IF mixers also suffer from port-to-port feedthroughs. Resembling a direct-conversion receiver, this section of the architecture follows the observations made for the topologies in Figs. 6.4 and 6.7.

The port-to-port feedthroughs of upconversion mixers are less critical, except for the LO-RF component. As explained in Chapter 4, the LO (or carrier) feedthrough corrupts the transmitted signal constellation and must be minimized.

6.1.2 Mixer Noise Figures

The noise figure of downconversion mixers is often a source of great confusion. For simplicity, let us consider a *noiseless* mixer with unity gain. As shown in Fig. 6.10, the spectrum sensed by the RF port consists of a signal component and the thermal noise of R_S in both the signal band and the image band. Upon downconversion, the signal, the noise in the signal band, and the noise in the image band are translated to ω_{IF}. Thus, the output SNR is *half* the input SNR if the two noise components have equal powers, i.e., the mixer exhibits a flat frequency response at its input from the image band to the signal band.

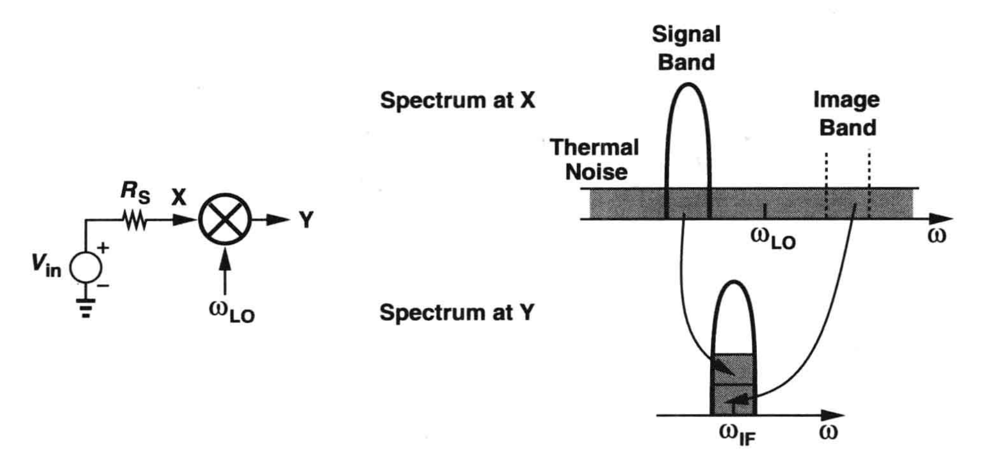


Figure 6.10 *SSB noise figure.*

We therefore say the noise figure of a noiseless mixer is 3 dB. This quantity is called the "single-sideband" (SSB) noise figure to indicate that the desired signal resides on only one side of the LO frequency, a common case in heterodyne receivers.

Now, consider the direct-conversion mixer shown in Fig. 6.11. In this case, only the noise in the signal band is translated to the baseband, thereby yielding equal input and output SNRs if the mixer is noiseless. The noise figure is thus equal to 0 dB. This quantity is called the "double-sideband" (DSB) noise figure to emphasize that the input signal resides on both sides of ω_{LO}, a common situation in direct-conversion receivers.

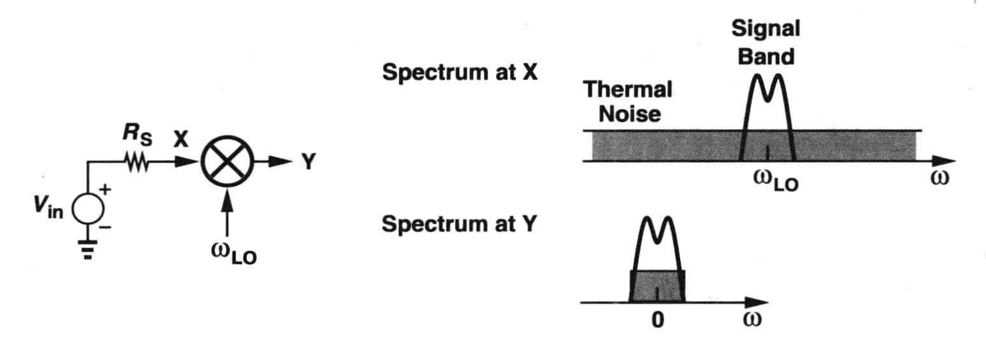


Figure 6.11 *DSB noise figure.*

In summary, the SSB noise figure of a mixer is 3 dB higher than its DSB noise figure if the signal and image bands experience equal gains at the RF port of the mixer. Typical noise figure meters measure the DSB NF and predict the SSB value by simply adding 3 dB.

Example 6.3

A student designs the heterodyne receiver of Fig. 6.12(a) for two cases: (1) ω_{LO1} is far from ω_{RF}; (2) ω_{LO1} lies *inside* the band and so does the image. Study the noise behavior of the receiver in the two cases.

Example 6.3 (*Continued*)

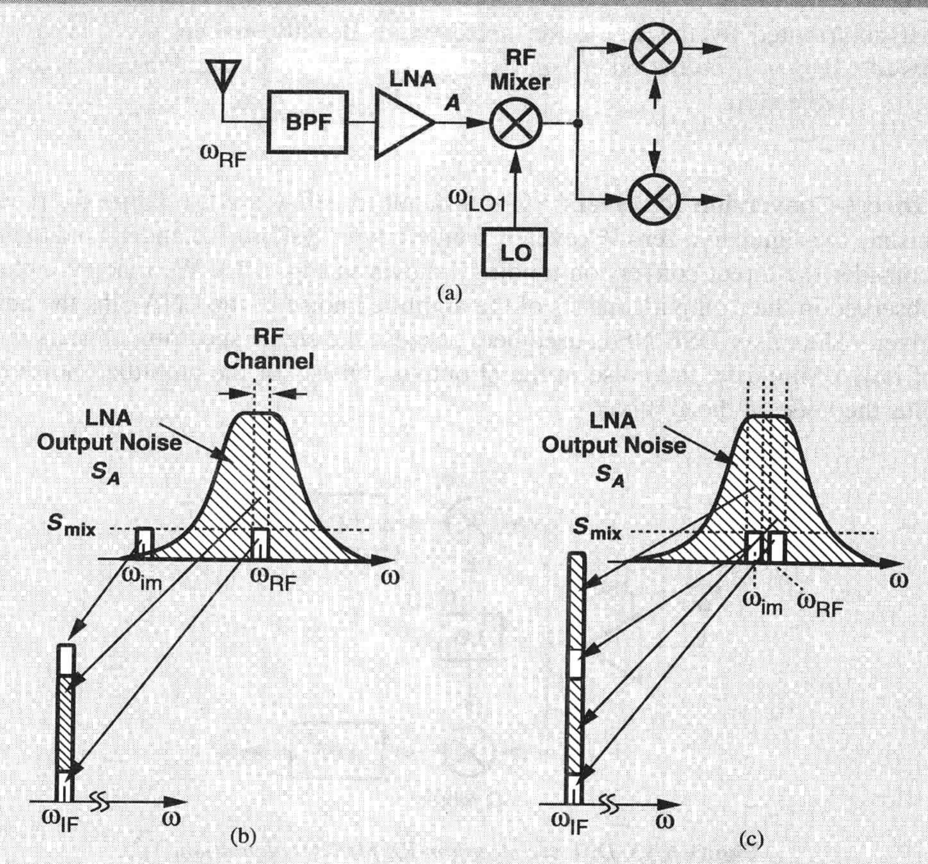


Figure 6.12 *(a) Heterodyne RX, (b) downconversion of noise with image located out of band, (c) downconversion of noise with image located in band.*

Solution:

In the first case, the selectivity of the antenna, the BPF, and the LNA suppresses the thermal noise in the image band. Of course, the RF mixer still folds its own noise. The overall behavior is illustrated in Fig. 6.12(b), where S_A denotes the noise spectrum at the output of the LNA and S_{mix} the noise in the input network of the mixer itself. Thus, the mixer downconverts three significant noise components to IF: the amplified noise of the antenna and the LNA around ω_{RF}, its own noise around ω_{RF}, and its image noise around ω_{im}.

In the second case, the noise produced by the antenna, the BPF, and the LNA exhibits a flat spectrum from the image frequency to the signal frequency. As shown in Fig. 6.12(c), the RF mixer now downconverts four significant noise components to IF: the output noise of the LNA around ω_{RF} and ω_{im}, and the input noise of the mixer around ω_{RF} and ω_{im}. We therefore conclude that the noise figure of the second frequency plan is substantially higher than that of the first. In fact, if the noise contributed by the mixer is much less

(Continues)

Example 6.3 *(Continued)*

than that contributed by the LNA, the noise figure penalty reaches 3 dB. The low-IF receivers of Chapter 4, on the other hand, do not suffer from this drawback because they employ image rejection.

NF of Direct-Conversion Receivers It is difficult to define a noise figure for receivers that translate the signal to a zero IF (even in a heterodyne system). To understand the issue, let us consider the direct-conversion topology shown in Fig. 6.13. We recognize that the noise observed in the I output consists of the amplified noise of the LNA plus the noise of the I mixer. (The mixer DSB NF is used here because the signal spectrum appears on both sides of ω_{LO}.) Similarly, the noise in the Q output consists of the amplified noise of the LNA plus the noise of the Q mixer.

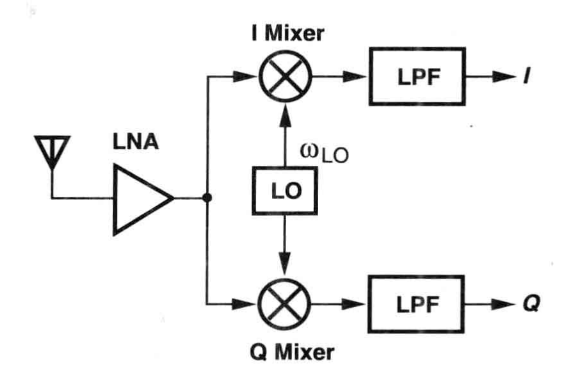


Figure 6.13 *Direct-conversion RX for NF calculation.*

But, how do we define the overall noise figure? Even though the system has *two* output ports, one may opt to define the NF with respect to only one,

$$\text{NF} = \frac{\text{SNR}_{in}}{\text{SNR}_I} = \frac{\text{SNR}_{in}}{\text{SNR}_Q}, \tag{6.5}$$

where SNR_I and SNR_Q denote the SNRs measured at the I and Q outputs, respectively. Indeed, this is the most common NF definition for direct-conversion receivers. However, since the I and Q outputs are eventually combined (possibly in the digital domain), the SNR in the final *combined* output would serve as a more accurate measure of the noise performance. Unfortunately, the manner in which the outputs are combined depends on the modulation scheme, thus making it difficult to obtain the output SNR. For example, as described in Chapter 4, an FSK receiver may simply sample the binary levels in the I output by the data edges in the Q output, leading to a *nonlinear* combining of the baseband quadrature signals. For these reasons, the NF is usually obtained according to Eq. (6.5), a somewhat pessimistic value because the signal component in the other output is ignored. Ultimately, the sensitivity of the receiver is characterized by the bit error rate, thereby avoiding the NF ambiguity.

Example 6.4

Consider the simple mixer shown in Fig. 6.14(a). Assuming $R_L \gg R_S$ and the LO has a 50% duty cycle, determine the output noise spectrum due to R_S, i.e., assume R_L is noiseless.

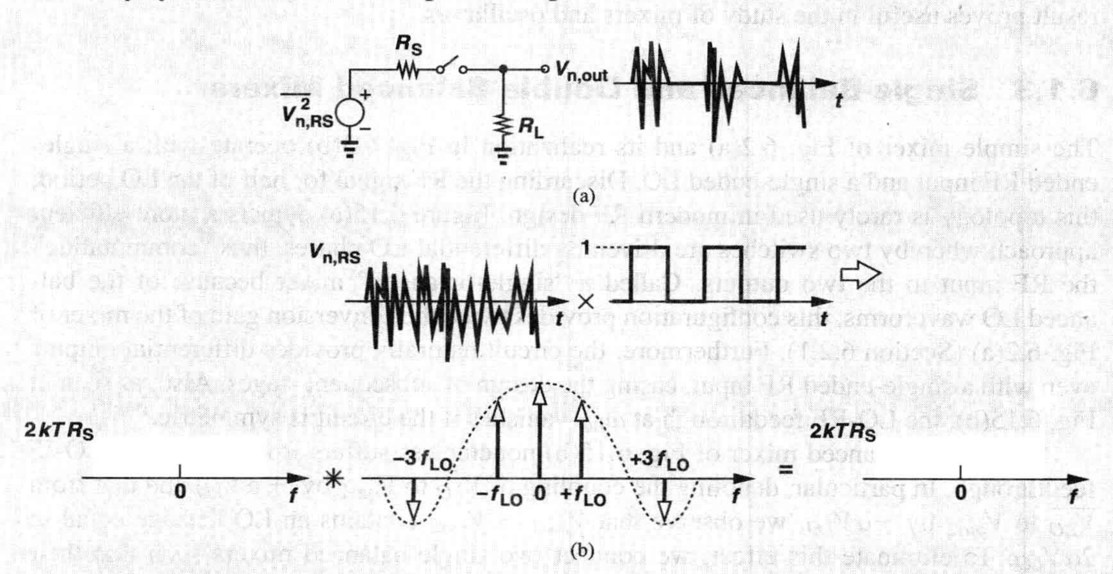


Figure 6.14 *(a) Passive mixer, (b) input and output signals in time and frequency domains.*

Solution:

Since V_{out} is equal to the noise of R_S for half of the LO cycle and equal to zero for the other half, we expect the output power density to be simply equal to half of that of the input, i.e., $\overline{V_{n,out}^2} = 2kTR_S$. (This is the *one-sided* spectrum.) To prove this conjecture, we view $V_{n,out}(t)$ as the product of $V_{n,RS}(t)$ and a square wave toggling between 0 and 1. The output spectrum is thus obtained by convolving the spectra of the two [Fig. 6.14(b)]. It is important to note that the *power* spectral density of the square wave has a sinc^2 envelope, exhibiting an impulse with an area of 0.5^2 at $f = 0$, two with an area of $(1/\pi)^2$ at $f = \pm f_{LO}$, etc. The output spectrum consists of (a) $2kTR_S \times 0.5^2$, (b) $2kTR_S$ shifted to the right and to the left by $\pm f_{LO}$ and multiplied by $(1/\pi)^2$, (c) $2kTR_S$ shifted to the right and to the left by $\pm 3f_{LO}$ and multiplied by $[1/(3\pi)]^2$, etc. We therefore write

$$\overline{V_{n,out}^2} = 2kTR_S\left[\frac{1}{2^2} + \frac{2}{\pi^2} + \frac{2}{(3\pi)^2} + \frac{2}{(5\pi)^2} + \cdots\right] \tag{6.6}$$

$$= 2kTR_S\left[\frac{1}{2^2} + \frac{2}{\pi^2}\left(1 + \frac{1}{3^2} + \frac{1}{5^2} + \cdots\right)\right]. \tag{6.7}$$

It can be proved that $1^{-2} + 3^{-2} + 5^{-2} + \cdots = \pi^2/8$. It follows that the *two-sided* output spectrum is equal to kTR_S and hence the one-sided spectrum is given by

$$\overline{V_{n,out}^2} = 2kTR_S. \tag{6.8}$$

The above example leads to an important conclusion: if white noise is switched on and off with 50% duty cycle, then the resulting spectrum is still white but carries half the power. More generally, if white noise is turned on for ΔT seconds and off for $T - \Delta T$ seconds, then the resulting spectrum is still white and its power is scaled by $\Delta T/T$. This result proves useful in the study of mixers and oscillators.

6.1.3 Single-Balanced and Double-Balanced Mixers

The simple mixer of Fig. 6.2(a) and its realization in Fig. 6.3(b) operate with a single-ended RF input and a single-ended LO. Discarding the RF signal for half of the LO period, this topology is rarely used in modern RF design. Figure 6.15(a) depicts a more efficient approach whereby two switches are driven by differential LO phases, thus "commutating" the RF input to the two outputs. Called a "single-balanced" mixer because of the balanced LO waveforms, this configuration provides twice the conversion gain of the mixer of Fig. 6.2(a) (Section 6.2.1). Furthermore, the circuit naturally provides differential outputs even with a single-ended RF input, easing the design of subsequent stages. Also, as seen in Fig. 6.15(b), the LO-RF feedthrough at ω_{LO} vanishes if the circuit is symmetric.[1]

The single-balanced mixer of Fig. 6.15(b) nonetheless suffers from significant LO-IF feedthrough. In particular, denoting the coupling of V_{LO} to V_{out1} by $+\alpha V_{LO}$ and that from $\overline{V_{LO}}$ to V_{out2} by $-\alpha V_{LO}$, we observe that $V_{out1} - V_{out2}$ contains an LO leakage equal to $2\alpha V_{LO}$. To eliminate this effect, we connect two single-balanced mixers such that their output LO feedthroughs cancel but their output signals do not. Shown in Fig. 6.16, such a topology introduces two opposing feedthroughs at each output, one from V_{LO} and another from $\overline{V_{LO}}$. The output signals remain intact because, when V_{LO} is high, $V_{out1} = V_{RF}^{+}$ and $V_{out2} = V_{RF}^{-}$, and when $\overline{V_{LO}}$ is high, $V_{out1} = V_{RF}^{-}$ and $V_{out2} = V_{RF}^{+}$. That is, $V_{out1} - V_{out2}$ is equal to $V_{RF}^{+} - V_{RF}^{-}$ for a high LO and $V_{RF}^{-} - V_{RF}^{+}$ for a low LO.

Called a "double-balanced" mixer, the circuit of Fig. 6.16 operates with both balanced LO waveforms and balanced RF inputs. It is possible to apply a single-ended RF input

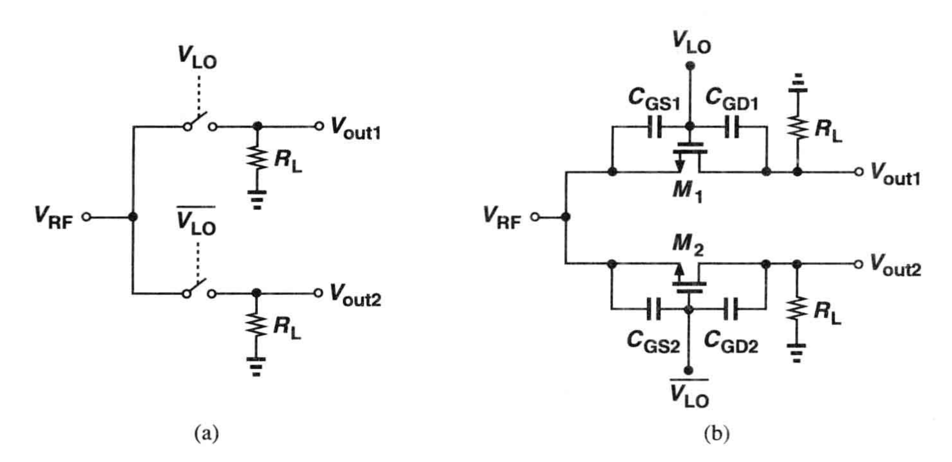


Figure 6.15 *(a) Single-balanced passive mixer, (b) implementation of (a).*

1. Due to nonlinearities, a component at $2\omega_{LO}$ still leaks to the input (Problem 6.3).

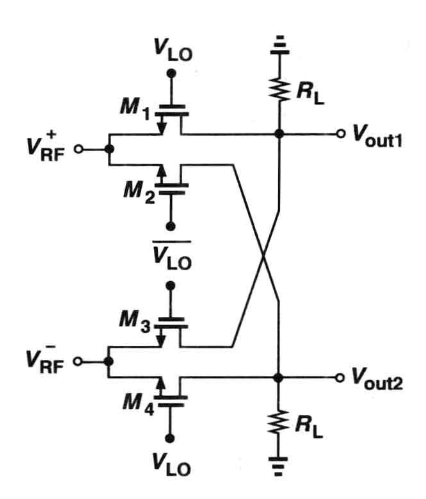


Figure 6.16 *Double-balanced passive mixer.*

(e.g., if the LNA is single-ended) while grounding the other, but at the cost of a higher input-referred noise.

Ideal LO Waveform What is the "ideal" LO waveform, a sinusoid or a square wave? Since each LO in an RF transceiver drives a mixer,[2] we note from the above observations that the LO waveform must ideally be a square wave to ensure abrupt switching and hence maximum conversion gain. For example, in the circuit of Fig. 6.16(b), if V_{LO} and $\overline{V_{LO}}$ vary gradually, then they remain approximately *equal* for a substantial fraction of the period (Fig. 6.17). During this time, all four transistors are on, treating V_{RF} as a *common-mode* input. That is, the input signal is "wasted" because it produces no differential component for roughly $2\Delta T$ seconds each period. As explained later, the gradual edges may also raise the noise figure.

At very high frequencies, the LO waveforms inevitably resemble sinusoids. We therefore choose a relatively large amplitude so as to obtain a high slew rate and ensure a minimum overlap time, ΔT.

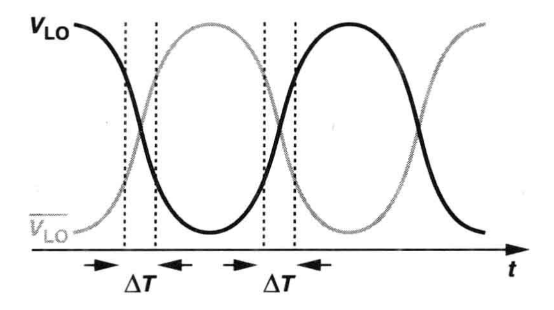


Figure 6.17 *LO waveforms showing when the switches are on simultaneously.*

2. One exception is when an LO drives only a frequency divider to avoid injection pulling (Chapter 4).

Since mixers equivalently multiply the RF input by a square wave, they can down-convert interferers located at the LO harmonics, a serious issue in broadband receiver. For example, an interferer at $3f_{LO}$ is attenuated by about only 10 dB as it appears in the baseband.

Passive and Active Mixers Mixers can be broadly categorized into "passive" and "active" topologies; each can be realized as a single-balanced or a double-balanced circuit. We study these types in the following sections.

6.2 PASSIVE DOWNCONVERSION MIXERS

The mixers illustrated in Figs. 6.15 and 6.16 exemplify passive topologies because their transistors do not operate as amplifying devices. We wish to determine the conversion gain, noise figure, and input impedance of a certain type of passive mixers. We first assume that the LO has a duty cycle of 50% and the RF input is driven by a voltage source.

6.2.1 Gain

Let us begin with Fig. 6.18(a) and note that the input is multiplied by a square wave toggling between 0 and 1. The first harmonic of this waveform has a peak amplitude of $2/\pi$ and can be expressed as $(2/\pi)\cos\omega_{LO}t$. In the frequency domain, this harmonic consists of two impulses at $\pm\omega_{LO}$, each having an area of $1/\pi$. Thus, as shown in Fig. 6.18(b), the convolution of an RF signal with these impulses creates the IF signal with a gain of $1/\pi$ (≈ -10 dB). The conversion gain is therefore equal to $1/\pi$ for abrupt LO switching. We call this topology a "return-to-zero" (RZ) mixer because the output falls to zero when the switch turns off.

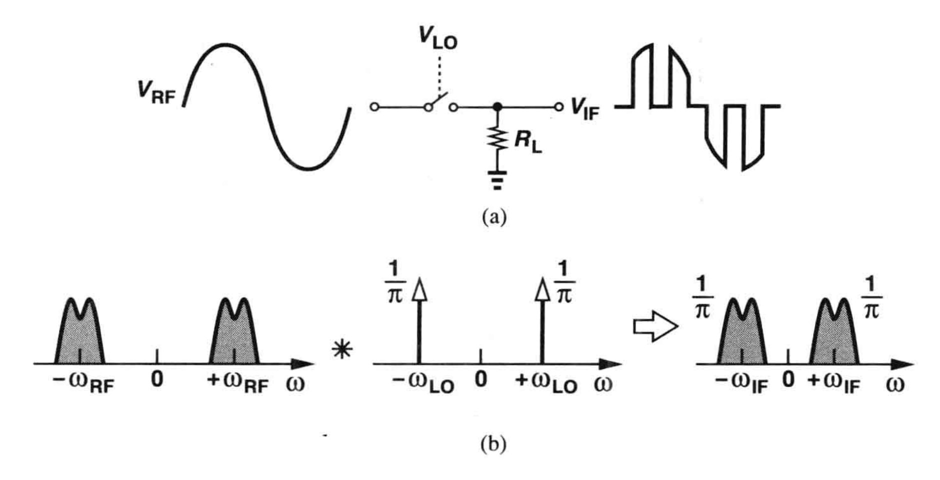


Figure 6.18 *(a) Input and output waveforms of a return-to-zero mixer, (b) corresponding spectra.*

Example 6.5

Explain why the mixer of Fig. 6.18 is ill-suited to direct-conversion receivers.

Solution:

Since the square wave toggling between 0 and 1 carries an average of 0.5, V_{RF} itself also appears at the output with a conversion gain of 0.5. Thus, low-frequency beat components resulting from even-order distortion in the preceding stage directly go to the output, yielding a low IP_2.

Example 6.6

Determine the conversion gain if the circuit of Fig. 6.18(a) is converted to a single-balanced topology.

Solution:

As illustrated in Fig. 6.19, the second output is similar to the first but shifted by 180°. Thus, the *differential* output contains twice the amplitude of each single-ended output. The conversion gain is therefore equal to $2/\pi$ (≈ -4 dB). Providing differential outputs and twice the gain, this circuit is superior to the single-ended topology of Fig. 6.18(a).

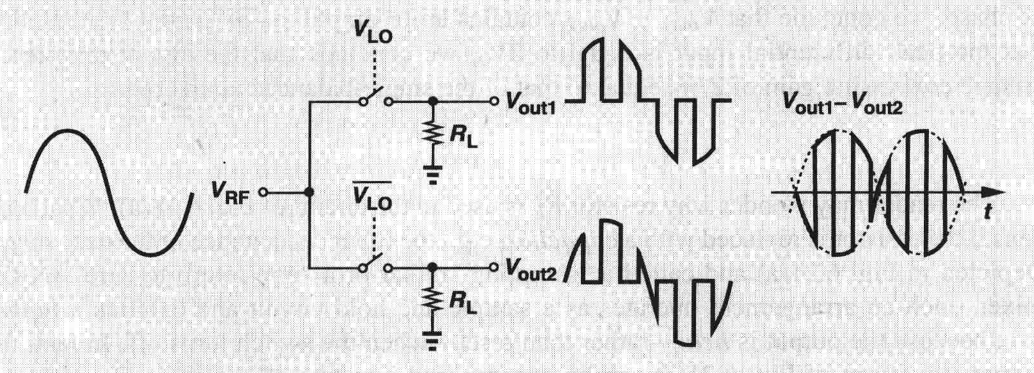


Figure 6.19 *Waveforms for passive mixer gain computation.*

Example 6.7

Determine the voltage conversion gain of a double-balanced version of the above topology [Fig. 6.20(a)]. (Decompose the differential output to return-to-zero waveforms.)

(Continues)

Example 6.7 *(Continued)*

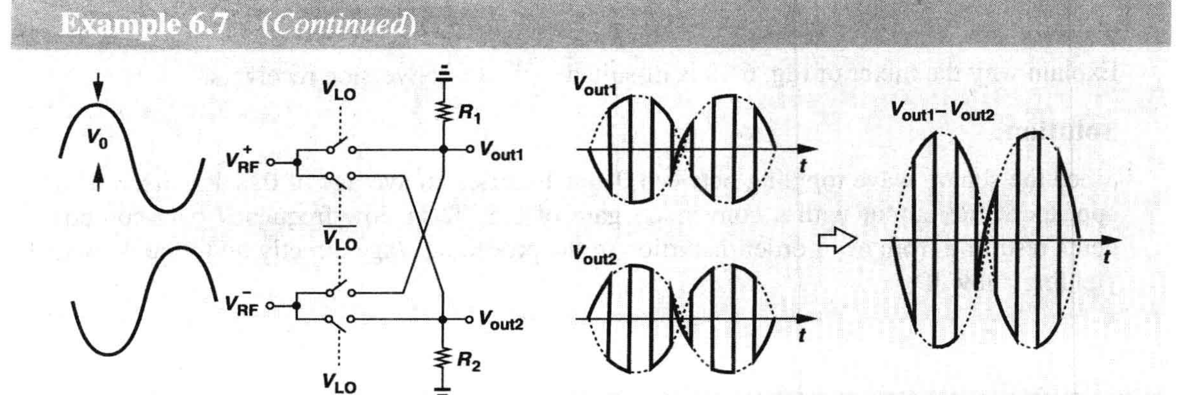


Figure 6.20 *(a) Double-balanced passive mixer, (b) output waveforms.*

Solution:

In this case, V_{out1} is equal to V_{RF}^+ for one half of the LO cycle and equal to V_{RF}^- for the other half, i.e., R_1 and R_2 can be omitted because the outputs do not "float." From the waveforms shown in Fig. 6.20(b), we observe that $V_{out1} - V_{out2}$ can be decomposed into two return-to-zero waveforms, each having a peak amplitude of $2V_0$ (why?). Since each of these waveforms generates an IF amplitude of $(1/\pi)2V_0$ and since the outputs are 180° out of phase, we conclude that $V_{out1} - V_{out2}$ contains an IF amplitude of $(1/\pi)(4V_0)$. Noting that the peak differential input is equal to $2V_0$, we conclude that the circuit provides a voltage conversion gain of $2/\pi$, equal to that of the single-balanced counterpart.

The reader may wonder why resistor R_L is used in the circuit of Fig. 6.18(a). What happens if the resistor is replaced with a *capacitor*, e.g., the input capacitance of the next stage? Depicted in Fig. 6.21(a) and called a "sampling" mixer or a "non-return-to-zero" (NRZ) mixer, such an arrangement operates as a sample-and-hold circuit and exhibits a *higher* gain because the output is *held*—rather than reset—when the switch turns off. In fact, the output waveform of Fig. 6.21(a) can be decomposed into two as shown in Fig. 6.21(b), where $y_1(t)$ is identical to the return-to-zero output in Fig. 6.18(a), and $y_2(t)$ denotes the additional output stored on the capacitor when S_1 is off. We wish to compute the voltage conversion gain.

We first recall the following Fourier transform pairs:

$$\sum_{k=-\infty}^{+\infty} \delta(t - kT) \leftrightarrow \frac{1}{T} \sum_{k=-\infty}^{+\infty} \delta\left(f - \frac{k}{T}\right) \tag{6.9}$$

$$x(t - T) \leftrightarrow e^{-j\omega T} X(f) \tag{6.10}$$

$$\prod\left(\frac{t}{T/2} - \frac{1}{2}\right) \leftrightarrow \frac{1}{j\omega}\left(1 - e^{-j\omega T/2}\right), \tag{6.11}$$

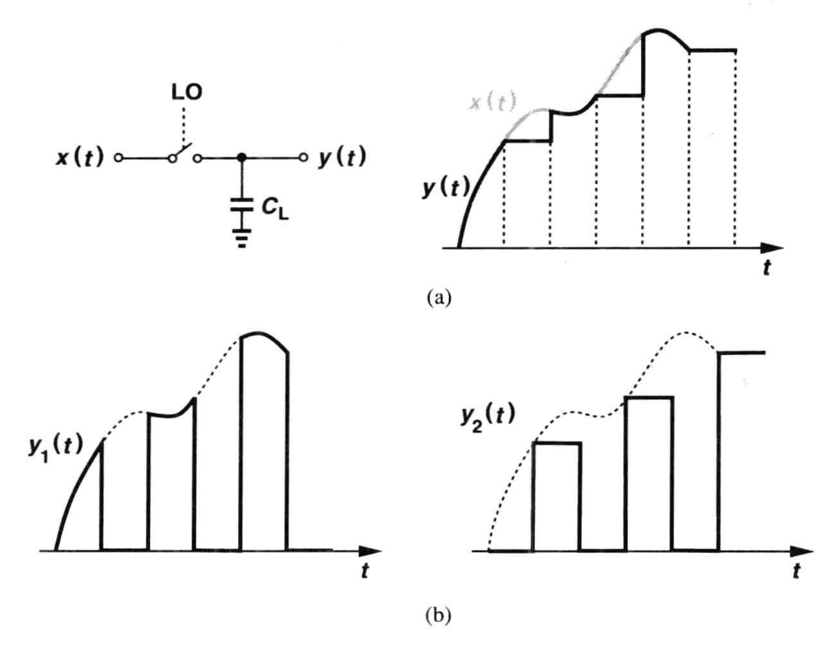


Figure 6.21 *(a) Sampling mixer, (b) output waveform decomposition.*

where $\prod[t/(T/2) - 1/2]$ represents a square pulse with an amplitude of 1 between $t = 0$ and $t = T/2$ and zero elsewhere. The right-hand side of Eq. (6.11) can also be expressed as a sinc. Since $y_1(t)$ is equal to $x(t)$ multiplied by a square wave toggling between zero and 1, and since such a square wave is equal to the convolution of a square pulse and a train of impulses [Fig. 6.22(a)], we have

$$y_1(t) = x(t)\left[\prod\left(\frac{t}{T_{LO}/2} - \frac{1}{2}\right) * \sum_{k=-\infty}^{+\infty} \delta(t - kT_{LO})\right], \tag{6.12}$$

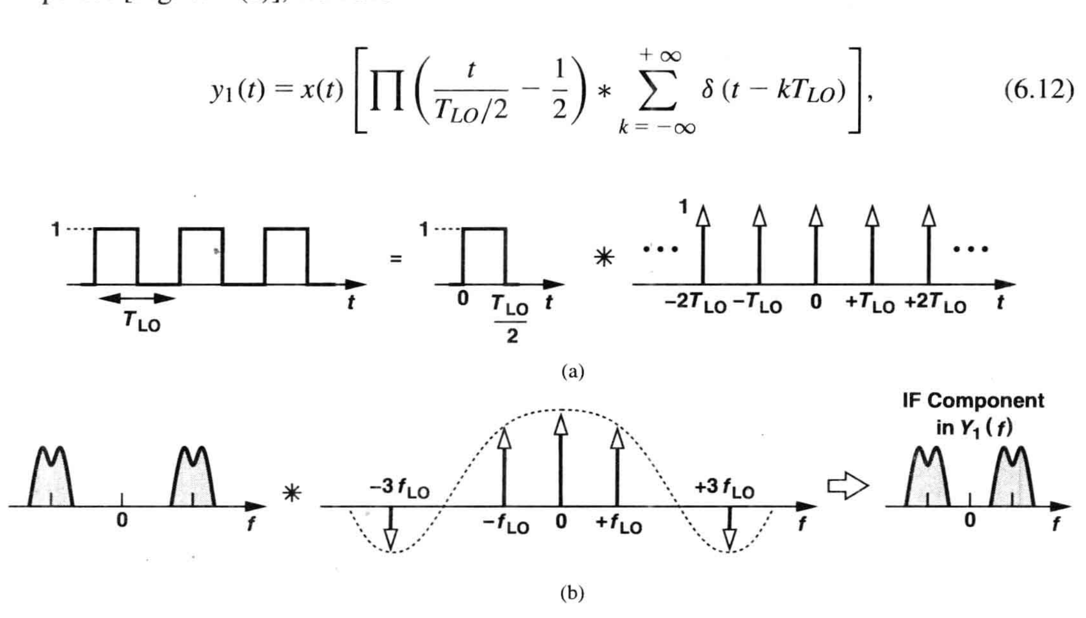


Figure 6.22 *(a) Decomposition of a square wave, (b) input and output spectra corresponding to $y_1(t)$.*

where T_{LO} denotes the LO period. It follows from Eqs. (6.9) and (6.11) that

$$Y_1(f) = X(f) * \left[\frac{1}{j\omega}\left(1 - e^{-j\omega T_{LO}/2}\right)\frac{1}{T_{LO}}\sum_{k=-\infty}^{+\infty}\delta\left(f - \frac{k}{T_{LO}}\right)\right]. \tag{6.13}$$

Figure 6.22(b) shows the corresponding spectra. The component of interest in $Y_1(f)$ lies at the IF and is obtained by setting k to ± 1:

$$Y_1(f)|_{IF} = X(f) * \left[\frac{1}{j\omega}\left(1 - e^{-j\omega T_{LO}/2}\right)\frac{1}{T_{LO}}\delta\left(f \pm \frac{1}{T_{LO}}\right)\right]. \tag{6.14}$$

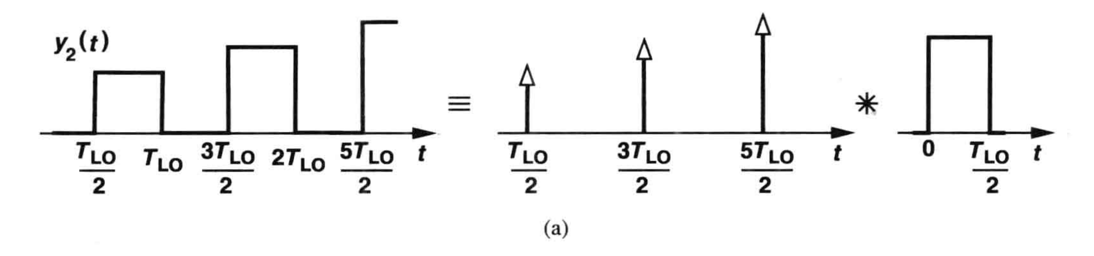
The impulse, in essence, computes $[1/(j\omega)][1 - \exp(-j\omega T_{LO}/2)]$ at $\pm 1/T_{LO}$, which amounts to $\pm T_{LO}/(j\pi)$. Multiplying this result by $(1/T_{LO})\delta(f \pm 1/T_{LO})$ and convolving it with $X(f)$, we have

$$Y_1(f)|_{IF} = \frac{X(f - f_{LO})}{j\pi} - \frac{X(f + f_{LO})}{j\pi}. \tag{6.15}$$

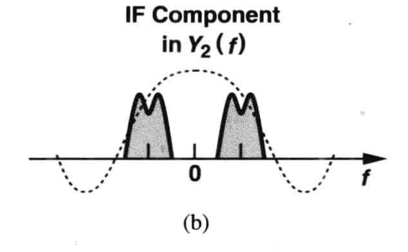
As expected, the conversion gain from $X(f)$ to $Y_1(f)$ is equal to $1/\pi$, but with a phase shift of 90°.

The second output in Fig. 6.21(b), $y_2(t)$, can be viewed as a train of impulses that sample the input and are subsequently convolved with a square pulse [Fig. 6.23(a)]. That is,

$$y_2(t) = \left[x(t)\sum_{k=-\infty}^{+\infty}\delta\left(t - kT_{LO} - \frac{T_{LO}}{2}\right)\right] * \prod\left(\frac{t}{T_{LO}/2} - \frac{1}{2}\right), \tag{6.16}$$



(a)



(b)

Figure 6.23 *(a) Decomposition of $y_2(t)$, (b) corresponding spectrum.*

and hence

$$Y_2(f) = \left[X(f) * \frac{1}{T_{LO}} \sum_{k=-\infty}^{+\infty} e^{-j\omega T_{LO}/2} \delta\left(f - \frac{k}{T_{LO}}\right)\right] \cdot \frac{1}{j\omega}\left(1 - e^{-j\omega T_{LO}/2}\right). \quad (6.17)$$

Figure 6.23(b) depicts the spectrum, revealing that shifted replicas of $X(f)$ are multiplied by a sinc envelope. Note the subtle difference between $Y_1(f)$ and $Y_2(f)$: in the former, each replica of $X(f)$ is simply scaled by a factor, whereas in the latter, each replica experiences a "droop" due to the sinc envelope. The component of interest in $Y_2(f)$ is obtained by setting k to ± 1:

$$Y_2(f)|_{IF} = \frac{1}{T_{LO}}[-X(f - f_{LO}) - X(f + f_{LO})]\left[\frac{1}{j\omega}\left(1 - e^{-j\omega T_{LO}/2}\right)\right]. \quad (6.18)$$

The term in the second set of square brackets must be calculated at the IF. If the IF is much lower than $2f_{LO}$, then $\exp(-j\omega_{IF} T_{LO}/2) \approx 1 - j\omega_{IF} T_{LO}/2$. Thus,

$$Y_2(f)|_{IF} \approx \frac{-X(f - f_{LO}) - X(f + f_{LO})}{2}. \quad (6.19)$$

Note that $Y_2(f)$ in fact contains a larger IF component than does $Y_1(f)$. The total IF output is therefore equal to

$$|Y_1(f) + Y_2(f)|_{IF} = \sqrt{\frac{1}{\pi^2} + \frac{1}{4}}[|X(f - f_{LO})| + |X(f + f_{LO})|] \quad (6.20)$$

$$= 0.593[|X(f - f_{LO})| + |X(f + f_{LO})|]. \quad (6.21)$$

If realized as a single-balanced topology (Fig. 6.24), the circuit provides a gain twice this value, $1.186 \approx 1.48$ dB. That is, a single-balanced sampling mixer exhibits about 5.5 dB higher gain than its return-to-zero counterpart. It is remarkable that, though a *passive* circuit, the single-ended sampling mixer actually has a voltage conversion gain greater than unity, and hence is a more attractive choice. The return-to-zero mixer is rarely used in modern RF design.

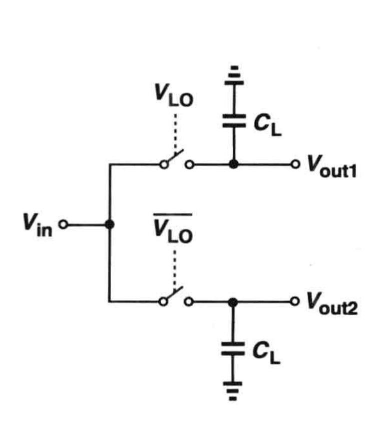


Figure 6.24 *Single-balanced sampling mixer.*

Example 6.8

Determine the voltage conversion gain of a double-balanced sampling mixer.

Solution:

Shown in Fig. 6.25, such a topology operates identically to the counterpart in Fig. 6.20(a). In other words, the capacitors play no role here because each output is equal to one of the inputs at any given point in time. The conversion gain is therefore equal to $2/\pi$, about 5.5 dB lower than that of the single-balanced topology of Fig. 6.24.

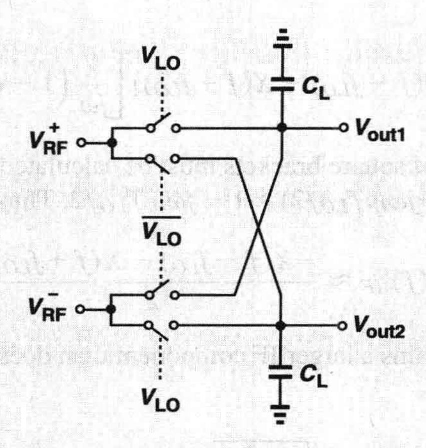


Figure 6.25 *Double-balanced sampling mixer.*

The above example may rule out the use of double-balanced sampling mixers. Since most receiver designs incorporate a single-ended LNA, this is not a serious limitation. However, if necessary, double-balanced operation can be realized through the use of two single-balanced mixers whose outputs are summed in the *current domain*. Illustrated conceptually in Fig. 6.26 [1], the idea is to retain the samples on the capacitors, convert each differential output voltage to a current by means of M_1–M_4, add their output currents, and

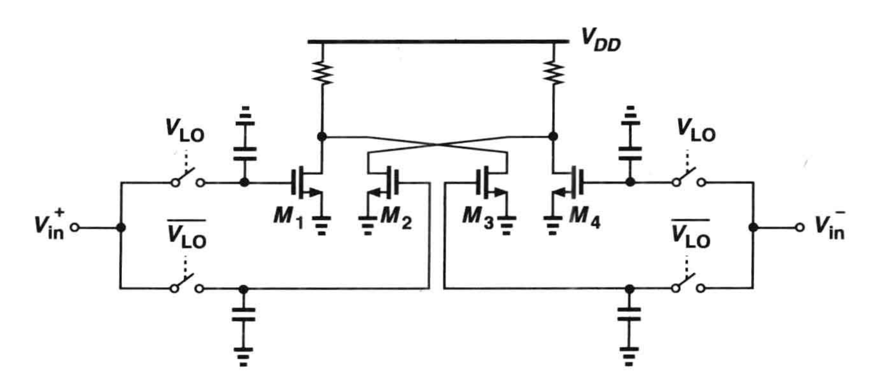


Figure 6.26 *Output combining of two single-balanced mixers in the current domain.*

apply the currents to load resistors, thus generating an output voltage. In this case, the mixer conversion gain is still equal to 1.48 dB.

6.2.2 LO Self-Mixing

Recall from Chapter 4 that the leakage of the LO waveform to the input of a mixer is added to the RF signal and mixed with the LO, generating a dc offset at the output. We now study this mechanism in the single-balanced sampling mixer. Consider the arrangement shown in Fig. 6.27(a), where R_S denotes the output impedance of the previous stage (the LNA). Suppose the LO waveforms and the transistors are perfectly symmetric. Then, due to the nonlinearity of C_{GS1} and C_{GS2} arising from large LO amplitudes, V_P does change with time but only at *twice* the LO frequency [Fig. 6.27(b)] (Problem 6.3). Upon mixing with the LO signal, this component is translated to f_{LO} and $3f_{LO}$—but *not* to dc. In other words, with perfectly-symmetric devices and LO waveforms, the mixer exhibits no LO self-mixing and hence no output dc offsets.

In practice, however, mismatches between M_1 and M_2 and within the oscillator circuit give rise to a finite LO leakage to node P. Accurate calculation of the resulting dc offset is difficult owing to the lack of data on various transistor, capacitor, and inductor mismatches that lead to asymmetries. A rough rule of thumb is 10–20 millivolts at the output of the mixer.

6.2.3 Noise

In this section, we study the noise behavior of return-to-zero and sampling mixers. Our approach is to determine the output noise spectrum, compute the output noise power in 1 Hz at the IF, and divide the result by the square of the conversion gain, thus obtaining the input-referred noise.

Let us begin with the RZ mixer, shown in Fig. 6.28. Here, R_{on} denotes the on-resistance of the switch. We assume a 50% duty cycle for the LO. The output noise is given by $4kT(R_{on}||R_L)$ when S_1 is on and by $4kTR_L$ when it is off. As shown in Example 6.4, on the

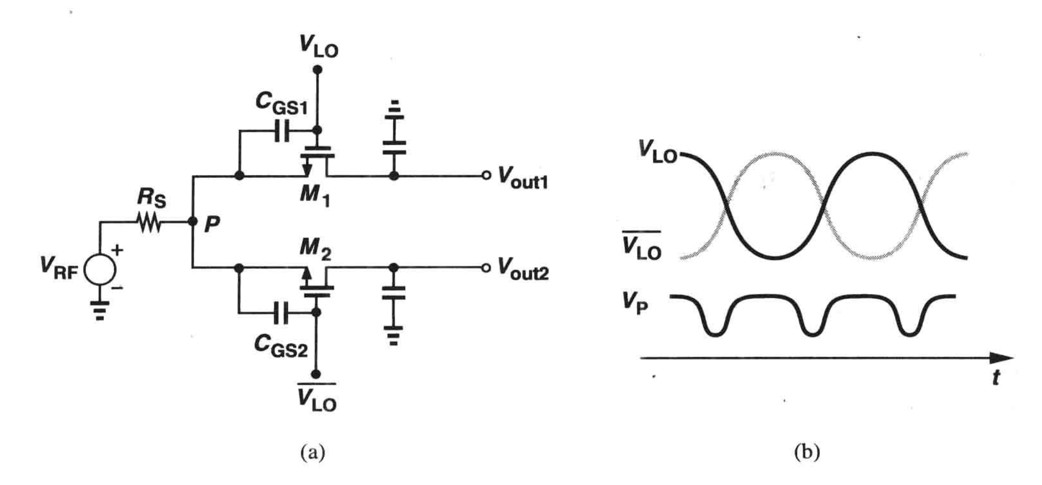


Figure 6.27 *(a) LO-RF leakage path in a sampling mixer, (b) LO and leakage waveforms.*

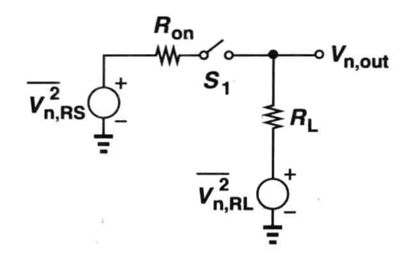


Figure 6.28 *RZ mixer for noise calculation.*

average, the output contains half of $4kT(R_{on}||R_L)$ and half of $4kTR_L$:

$$\overline{V^2_{n,out}} = 2kT[(R_{on}||R_L) + R_L]. \tag{6.22}$$

If we select $R_{on} \ll R_L$ so as to minimize the conversion loss, then

$$\overline{V^2_{n,out}} \approx 2kTR_L. \tag{6.23}$$

Dividing this result by $1/\pi^2$, we have

$$\overline{V^2_{n,in}} \approx 2\pi^2 kTR_L \tag{6.24}$$

$$\approx 20kTR_L. \tag{6.25}$$

That is, the noise power of R_L ($= 4kTR_L$) is "amplified" by a factor of 5 when referred to the input.

Example 6.9

If $R_{on} = 100\ \Omega$ and $R_L = 1\ \text{k}\Omega$, determine the input-referred noise of the above RZ mixer.

Solution:

We have

$$\sqrt{\overline{V^2_{n,in}}} = 8.14\,\text{nV}/\sqrt{\text{Hz}}. \tag{6.26}$$

This noise would correspond to a noise figure of $10\log[1 + (8.14/0.91)^2] = 19\,\text{dB}$ in a 50-Ω system.

The reader may wonder if our choice $R_{on} \ll R_L$ is optimum. If R_L is very high, the output noise decreases but so does the conversion gain. We now remove the assumption $R_{on} \ll R_L$ and express the voltage conversion gain as $(1/\pi)R_L/(R_{on} + R_L)$. Dividing Eq. (6.22) by the square of this value gives

$$\overline{V^2_{n,in}} = 2\pi^2 kT\frac{(R_{on} + R_L)(2R_{on} + R_L)}{R_L}. \tag{6.27}$$

This function reaches a minimum of

$$\overline{V^2_{n,in,min}} = 2\pi^2(2\sqrt{2}+3)kTR_{on} \quad (6.28)$$

$$\approx 117kTR_{on} \quad (6.29)$$

for $R_L = \sqrt{2}R_{on}$. For example, if $R_{on} = 100\,\Omega$ and $R_L = \sqrt{2} \times 100\,\Omega$, then the input-referred noise voltage is equal to $6.96\,\text{nV}/\sqrt{\text{Hz}}$ (equivalent to an NF of 17.7 dB in a 50-Ω system).

In reality, the output noise voltages calculated above are pessimistic because the input capacitance of the following stage limits the noise bandwidth, i.e., the noise is no longer white. This point becomes clearer in our study of the sampling mixer.

We now wish to compute the output noise spectrum of a sampling mixer. The output noise at the IF can then be divided by the conversion gain to obtain the input-referred noise voltage. We begin with three observations. First, in the simple circuit of Fig. 6.29(a) (where R_1 denotes the switch resistance), if $V_{in} = 0$,

$$\overline{V^2_{n,LPF}} = \overline{V^2_{nR1}} \frac{1}{1 + (R_1 C_1 \omega)^2}, \quad (6.30)$$

where $\overline{V^2_{nR1}} = 2kTR_1$ (for $-\infty < \omega < +\infty$). We say the noise is "shaped" by the filter.[3] Second, in the switching circuit of Fig. 6.29(b), the output is equal to the shaped noise of

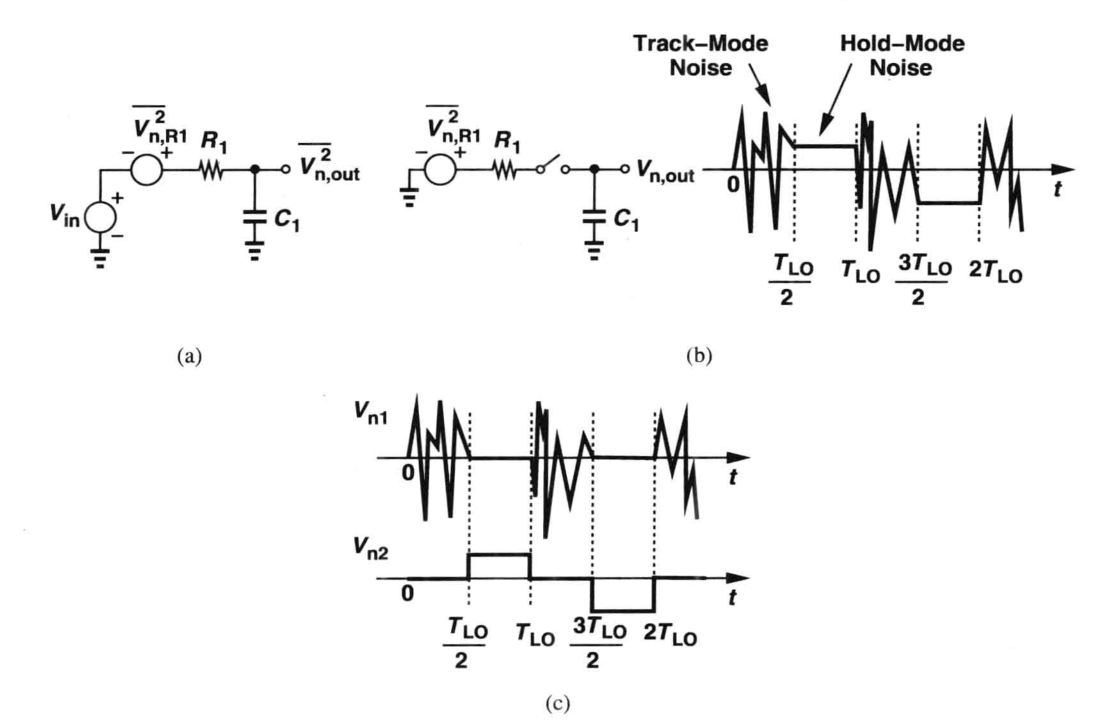


Figure 6.29 *(a) Equivalent circuit of sampling mixer for noise calculations, (b) noise in on and off states, and (c) decomposition of output waveform.*

3. Recall from basic analog circuits that the integral of this output noise from 0 to ∞ is equal to kT/C_1.

R_1 when S_1 is on and a *sampled*, constant value when it is off. Third, in a manner similar to the gain calculation in Fig. 6.21, we can decompose the output into two waveforms V_{n1} and V_{n2} as shown in Fig. 6.29(c).

It is tempting to consider the overall output spectrum as the sum of the spectra of V_{n1} and V_{n2}. However, as explained below, the low-frequency noise components generated by R_1 create *correlation* between the track-mode and hold-mode noise waveforms. For this reason, we proceed as follows: (1) compute the spectrum of V_{n1} while excluding the low-frequency components in the noise of R_1, (2) do the same for V_{n2}, and (3) add the contribution of the low-frequency components to the final result. In the derivations below, we refer to the first two as simply the spectra of V_{n1} and V_{n2} even though $V_{n1}(t)$ and $V_{n2}(t)$ in Fig. 6.29 are affected by the low-frequency noise of R_1. Similarly, we use the notation $\overline{V_{n,LPF}^2}(f)$ even though its low-frequency components are removed and considered separately.

Spectrum of V_{n1} To calculate the spectrum of V_{n1}, we view this waveform as the product of $V_{n,LPF}(t)$ and a square wave toggling between 0 and 1. As shown in Fig. 6.30, the spectrum of V_{n1} is given by the convolution of $\overline{V_{n,LPF}^2}(f)$ and the power spectral density of the square wave (impulses with a sinc^2 envelope). In practice, the sampling bandwidth of the mixer, $1/(R_1C_1)$, rarely exceeds $3\omega_{LO}$, and hence

$$\overline{V_{n1}^2}(f) = 2 \times \left(\frac{1}{\pi^2} + \frac{1}{9\pi^2}\right) \frac{2kTR_1}{1 + (2\pi R_1 C_1 f)^2}, \tag{6.31}$$

where the factor of 2 on the right-hand side accounts for the aliasing of components at negative and positive frequencies. At low output frequencies, this expression reduces to

$$\overline{V_{n1}^2} = 0.226(2kTR_1). \tag{6.32}$$

Note that this is the two-sided spectrum of $\overline{V_{n1}^2}$.

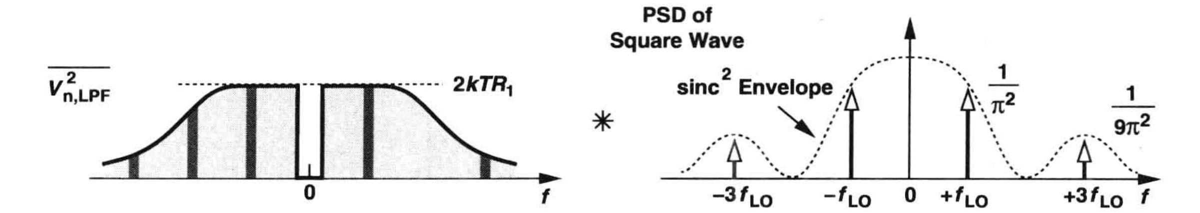


Figure 6.30 *Aliasing in V_{n1}.*

Spectrum of V_{n2} The spectrum of V_{n2} in Fig. 6.29(c) can be obtained using the approach illustrated in Fig. 6.21 for the conversion gain. That is, V_{n2} is equivalent to sampling $V_{n,LPF}$ by a train of impulses and convolving the result with a square pulse, $\prod[t/(2T_{LO}) - 1/2]$. We must therefore convolve the spectrum of $V_{n,LPF}$ with a train of impulses (each having an area of $1/T_{LO}^2$) and multiply the result by a sinc^2 envelope. As shown in Fig. 6.31, the

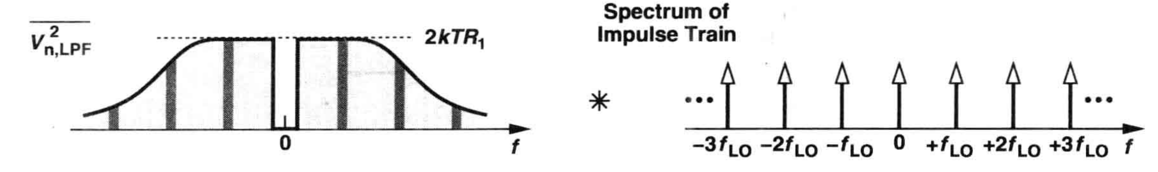


Figure 6.31 *Aliasing in V_{n2}.*

convolution translates noise components around $\pm f_{LO}$, $\pm 2f_{LO}$, etc., to the IF. The sum of these aliased components is given by

$$\overline{V^2_{n,\text{alias}}} = 2 \times \frac{2kTR_1}{T^2_{LO}} \left[\frac{1}{1 + 4\pi^2 R_1^2 C_1^2 f_{LO}^2} + \frac{1}{1 + 4\pi^2 R_1^2 C_1^2 \left(2f_{LO}^2\right)} + \cdots \right] \tag{6.33}$$

$$= 2 \times \frac{2kTR_1}{T^2_{LO}} \sum_{n=1}^{\infty} \frac{1}{1 + a^2 n^2}, \tag{6.34}$$

where $a = 2\pi R_1 C_1 f_{LO}$. For the summation in Eq. (6.34), we have

$$\sum_{n=1}^{\infty} \frac{1}{1 + a^2 n^2} = \frac{1}{2} \left(\frac{\pi}{a} \coth \frac{\pi}{a} - 1 \right), \tag{6.35}$$

Also, typically $(2\pi R_1 C_1)^{-1} > f_{LO}$ and hence $\coth(2R_1 C_1 f_{LO})^{-1} \approx 1$. It follows that

$$\overline{V^2_{n,\text{alias}}} = \frac{kT}{T^2_{LO}} \left(\frac{1}{C_1 f_{LO}} - 2R_1 \right). \tag{6.36}$$

This result must be multiplied by the sinc^2 envelope, $|(j\omega)^{-1}[1 - \exp(-j\omega T_{LO}/2)]|^2$, which has a magnitude of $T^2_{LO}/4$ at low frequencies. Thus, the two-sided IF spectrum of V_{n2} is given by

$$\overline{V^2_{n2}} = kT \left(\frac{1}{4C_1 f_{LO}} - \frac{R_1}{2} \right). \tag{6.37}$$

Correlation Between V_{n1} and V_{n2} We must now consider the correlation between V_{n1} and V_{n2} in Fig. 6.29. The correlation arises from two mechanisms: (1) as the circuit enters the track mode, the previous sampled value takes a finite time to vanish, and (2) when the circuit enters the hold mode, the frozen noise value, V_{n2}, is partially correlated with V_{n1}. The former mechanism is typically negligible because of the short track time constant. For the latter, we recognize that the noise frequency components far below f_{LO} remain relatively *constant* during the track and hold modes (Fig. 6.32); it is as if they experienced a zero-order hold operation and hence a conversion gain of unity. Thus, the R_1 noise components from 0 to roughly $f_{LO}/10$ directly appear at the output, adding a noise PSD of $2kTR_1$.

Summing the *one-sided* spectra of V_{n1} and V_{n2} and the low-frequency contribution, $4kTR_1$, gives the total (one-sided) output noise at the IF:

$$\overline{V^2_{n,out,IF}} = kT \left(3.9R_1 + \frac{1}{2C_1 f_{LO}} \right). \tag{6.38}$$

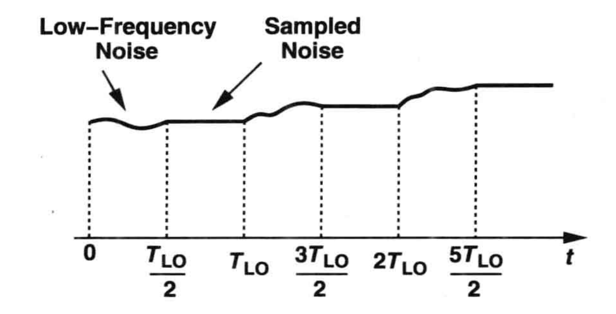


Figure 6.32 *Correlation between noise components in acquisition and hold modes.*

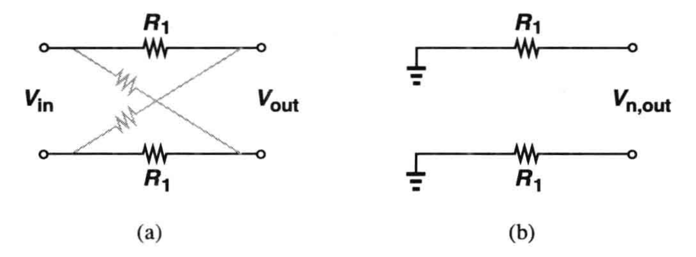


Figure 6.33 *(a) Equivalent circuit of double-balanced passive mixer, (b) simplified circuit.*

The input-referred noise is obtained by dividing this result by $1/\pi^2 + 1/4$:

$$\overline{V^2_{n,in}} = 2.85kT\left(3.9R_1 + \frac{1}{2C_1 f_{LO}}\right). \tag{6.39}$$

Note that [2] and [3] do not predict the dependence on R_1 or C_1.

For a single-balanced topology, the differential output exhibits a noise power twice that given by Eq. (6.38), but the *voltage* conversion gain is twice as high. Thus, the input-referred noise of a single-balanced passive (sampling) mixer is equal to

$$\overline{V^2_{n,in,SB}} = \frac{kT}{2\left(\frac{1}{\pi^2} + \frac{1}{4}\right)}\left(3.9R_1 + \frac{1}{2C_1 f_{LO}}\right) \tag{6.40}$$

$$= 1.42kT\left(3.9R_1 + \frac{1}{2C_1 f_{LO}}\right). \tag{6.41}$$

Let us now study the noise of a double-balanced passive mixer. As mentioned in Example 6.8, the behavior of the circuit does not depend much on the absence or presence of load capacitors. With abrupt LO edges, a resistance equal to R_1 appears between one input and one output at any point in time [Fig. 6.33(a)]. Thus, from Fig. 6.33(b), $\overline{V^2_{n,out}} = 8kTR_1$. Since the voltage conversion is equal to $2/\pi$,

$$\overline{V^2_{n,in}} = 2\pi^2 kTR_1. \tag{6.42}$$

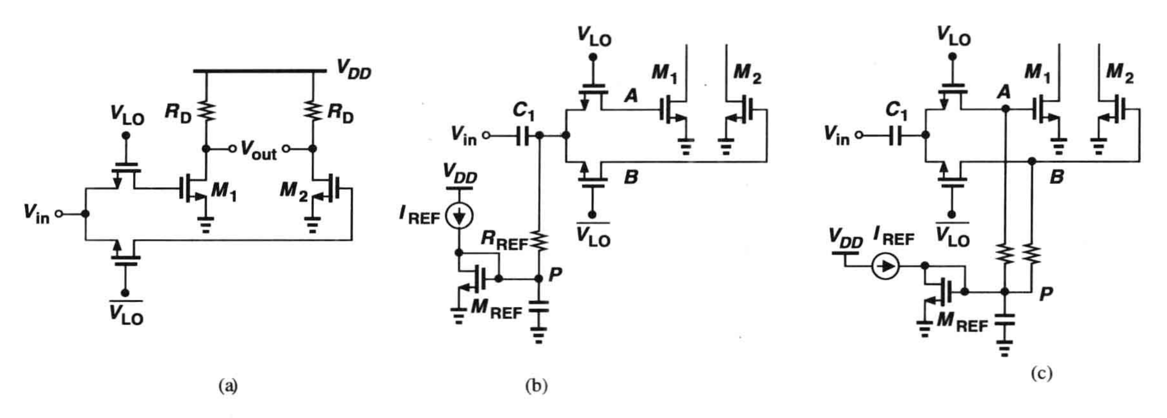


Figure 6.34 *(a) Passive mixer followed by gain stage, (b) bias path at the RF input, (c) bias path at the baseband output.*

The low gain of passive mixers makes the noise of the *subsequent* stage critical. Figure 6.34(a) shows a typical arrangement, where a quasi-differential pair (Chapter 5) serves as an amplifier and its input capacitance holds the output of the mixer. Each common-source stage exhibits an input-referred noise voltage of

$$\overline{V_{n,CS}^2} = \frac{4kT\gamma}{g_m} + \frac{4kT}{g_m^2 R_D}. \tag{6.43}$$

This power should be doubled to account for the two halves of the circuit and added to the mixer output noise power.

How is the circuit of Fig. 6.34(a) biased? Depicted in Fig. 6.34(b) is an example. Here, the bias of the preceding stage (the LNA) is blocked by C_1, and the network consisting of R_{REF}, M_{REF}, and I_{REF} defines the bias current of M_1 and M_2. As explained in Chapter 5, resistor R_{REF} is chosen much greater than the output resistance of the preceding stage. We typically select $W_{REF} \approx 0.2W_{1,2}$ so that $I_{D1,2} \approx 5I_{REF}$.

In the circuit of Fig. 6.34(b), the dc voltages at nodes A and B are equal to V_P unless LO self-mixing produces a dc offset between these two nodes. The reader may wonder if the circuit can be rearranged as shown in Fig. 6.34(c) so that the bias resistors provide a path to *remove* the dc offset. The following example elaborates on this point.

Example 6.10

A student considers the arrangement shown in Fig. 6.35(a), where V_{in} models the LO leakage to the input. The student then decides that the arrangement in Fig. 6.35(b) is *free* from dc offsets, reasoning that a positive dc voltage, V_{dc}, at the output would lead to a dc current, V_{dc}/R_L, through R_L and hence an equal current through R_S. This is impossible because it gives rise to a *negative* voltage at node X. Does the student deserve an A?

(Continues)

Example 6.10 *(Continued)*

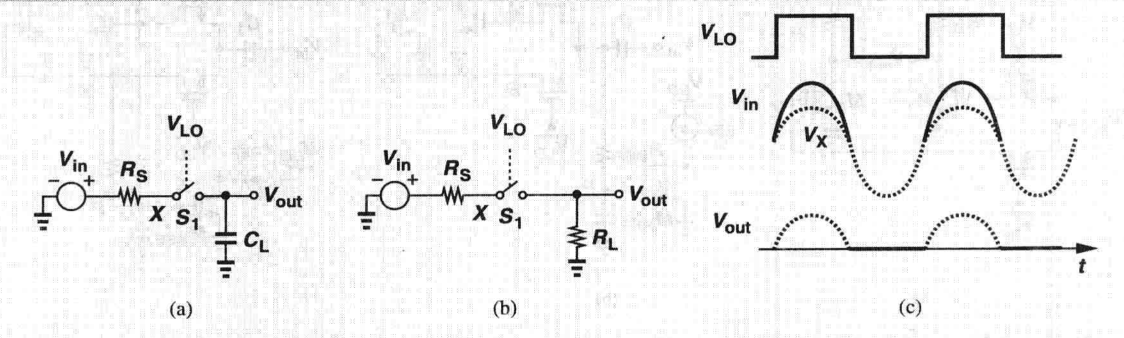


Figure 6.35 *(a) Sampling and (b) RZ mixer, (c) RZ mixer waveforms.*

Solution:

The average voltage at node X *can* be negative. As shown in Fig. 6.35(c), V_X is an attenuated version of V_{in} when S_1 is on and equal to V_{in} when S_1 is off. Thus, the average value of V_X is negative while R_L carries a finite average current as well. That is, the circuit of Fig. 6.35(b) still suffers from a dc offset.

6.2.4 Input Impedance

Passive mixers tend to present an appreciable load to LNAs. We therefore wish to formulate the input impedance of passive sampling mixers.

Consider the circuit depicted in Fig. 6.36, where S_1 is assumed ideal for now. Recall from Fig. 6.21 that the output voltage can be viewed as the sum of two waveforms $y_1(t)$ and $y_2(t)$, given by Eqs. (6.12) and (6.16), respectively. The current drawn by C_1 in Fig. 6.36 is equal to

$$i_{out}(t) = C_1 \frac{dy}{dt}. \tag{6.44}$$

Moreover, $i_{in}(t) = i_{out}(t)$. Taking the Fourier transform, we thus have

$$I_{in}(f) = C_1 j\omega Y(f), \tag{6.45}$$

where $Y(f)$ is equal to the sum of $Y_1(f)$ and $Y_2(f)$.

As evident from Figs. 6.22 and 6.23, $Y(f)$ contains many frequency components. We must therefore reflect on the meaning of the "input impedance." Since the input voltage signal, $x(t)$, is typically confined to a narrow bandwidth, we seek frequency components in $I_{in}(f)$ that lie within the bandwidth of $x(t)$. To this end, we set k in Eqs. (6.13) and (6.17) to *zero* so that $X(f)$ is simply convolved with $\delta(f)$ [i.e., the center frequency of $X(f)$ does

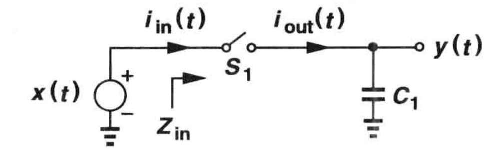


Figure 6.36 *Input impedance of sampling mixer.*

not change]. (This stands in contrast to gain and noise calculations, where k was chosen to translate $X(f)$ to the IF of interest.) It follows that

$$\frac{I_{in}(f)}{C_1 j\omega} = X(f) * \left[\frac{1}{j\omega}\left(1 - e^{-j\omega T_{LO}/2}\right)\frac{1}{T_{LO}}\delta(f)\right] + \left\{X(f) * \left[\frac{1}{T_{LO}}e^{-j\omega T_{LO}/2}\delta(f)\right]\right\}\frac{1}{j\omega}\left(1 - e^{-j\omega T_{LO}/2}\right). \tag{6.46}$$

In the square brackets in the first term, ω must be set to zero to evaluate the impulse at $f = 0$. Thus, the first term reduces to $(1/2)X(f)$. In the second term, the exponential in the square brackets must also be calculated at $\omega = 0$. Consequently, the second term simplifies to $(1/T_{LO})X(f)[1/(j\omega)][1 - \exp(-j\omega T_{LO}/2)]$. We then arrive at an expression for the input admittance:

$$\frac{I_{in}(f)}{X(f)} = jC_1\omega\left[\frac{1}{2} + \frac{1}{j\omega T_{LO}}\left(1 - e^{-j\omega T_{LO}/2}\right)\right]. \tag{6.47}$$

Note that the on-resistance of the switch simply appears in series with the inverse of (6.47).

It is instructive to examine Eq. (6.47) for a few special cases. If ω (the input frequency) is much less than ω_{LO}, then the second term in the square brackets reduces to $1/2$ and

$$\frac{I_{in}(f)}{X(f)} = jC_1\omega. \tag{6.48}$$

In other words, the entire capacitance is seen at the input [Fig. 6.37(a)]. If $\omega \approx 2\pi f_{LO}$ (as in direct-conversion receivers), then the second term is equal to $1/(j\pi)$ and

$$\frac{I_{in}(f)}{X(f)} = \frac{jC_1\omega}{2} + 2fC_1. \tag{6.49}$$

The input impedance thus contains a parallel resistive component equal to $1/(2fC_1)$ [Fig. 6.37(b)]. Finally, if $\omega \gg 2\pi f_{LO}$, the second term is much less than the first, yielding

$$\frac{I_{in}(f)}{X(f)} = \frac{jC_1\omega}{2}. \tag{6.50}$$

For the input impedance of a single-balanced mixer, we must add the switch on-resistance, R_1, to the inverse of Eq. (6.47) and halve the result. If $\omega \approx \omega_{LO}$, then

$$Z_{in,SB} = \frac{1}{2}\left[R_1 + \frac{1}{\dfrac{jC_1\omega}{2} + 2fC_1}\right]. \tag{6.51}$$

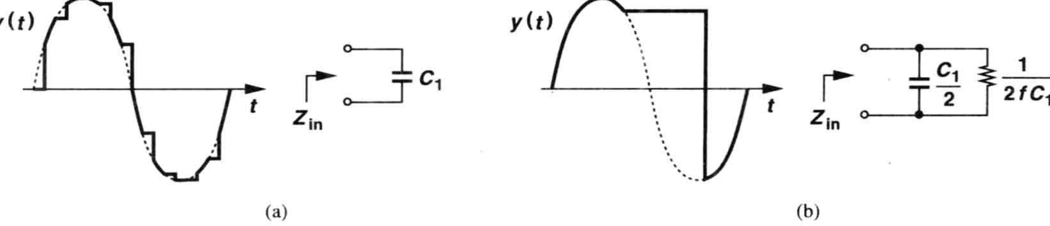


Figure 6.37 *Input impedance of passive mixer for (a) $\omega \ll \omega_{LO}$ and (b) $\omega \approx \omega_{LO}$.*

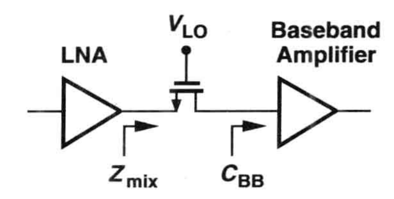


Figure 6.38 *Baseband input capacitance reflected at the input of passive mixer.*

Flicker Noise An important advantage of passive mixers over their active counterparts is their much lower output flicker noise. This property proves critical in narrowband applications, where $1/f$ noise in the baseband can substantially corrupt the downconverted channel.

MOSFETs produce little flicker noise if they carry a small current [4], a condition satisfied in a passive sampling mixer if the load capacitance is relatively small. However, the low gain of passive mixers makes the $1/f$ noise contribution of the *subsequent* stage critical. Thus, the baseband amplifier following the mixer must employ large transistors, presenting a large load capacitance to the mixer (Fig. 6.38). As explained above, C_{BB} manifests itself in the input impedance of the mixer, Z_{mix}, thereby loading the LNA.

LO Swing Passive MOS mixers require large (rail-to-rail) LO swings, a disadvantage with respect to active mixers. Since LC oscillators typically generate large swings, this is not a serious drawback, at least at moderate frequencies (up to 5 or 10 GHz).

In Chapter 13, we present the design of a passive mixer followed by a baseband amplifier for 11a/g applications.

6.2.5 Current-Driven Passive Mixers

The gain, noise, and input impedance analyses carried out in the previous sections have assumed that the RF input of passive mixers is driven by a voltage source. If driven by a current source, such mixers exhibit different properties. Figure 6.39(a) shows a conceptual arrangement where the LNA has a relatively high output impedance, approximating a current source. The passive mixer still carries no bias current so as to achieve low flicker noise and it drives a general impedance Z_{BB}. Voltage-driven and current-driven passive mixers entail a number of interesting differences.

First, the input impedance of the current-driven mixer in Fig. 6.39 is quite different from that of the voltage-driven counterpart. The reader may find this strange. Indeed, familiar circuits exhibit an input impedance that is independent of the source impedance: we can calculate the input impedance of an LNA by applying a voltage or a current source to the input port. A passive mixer, on the other hand, does not satisfy this intuition because it is a *time-variant* circuit. To determine the input impedance of a current-driven single-balanced mixer, we consider the simplified case depicted in Fig. 6.39(b), where the on-resistance of the switches is neglected. We wish to calculate $Z_{in}(f) = V_{RF}(f)/I_{in}(f)$ in the vicinity of the carrier (LO) frequency, assuming a 50% duty cycle for the LO.

The input current is routed to the upper arm for 50% of the time and flows through Z_{BB}. In the time domain [5],

$$V_1(t) = [i_{in}(t) \times S(t)] * h(t), \tag{6.52}$$

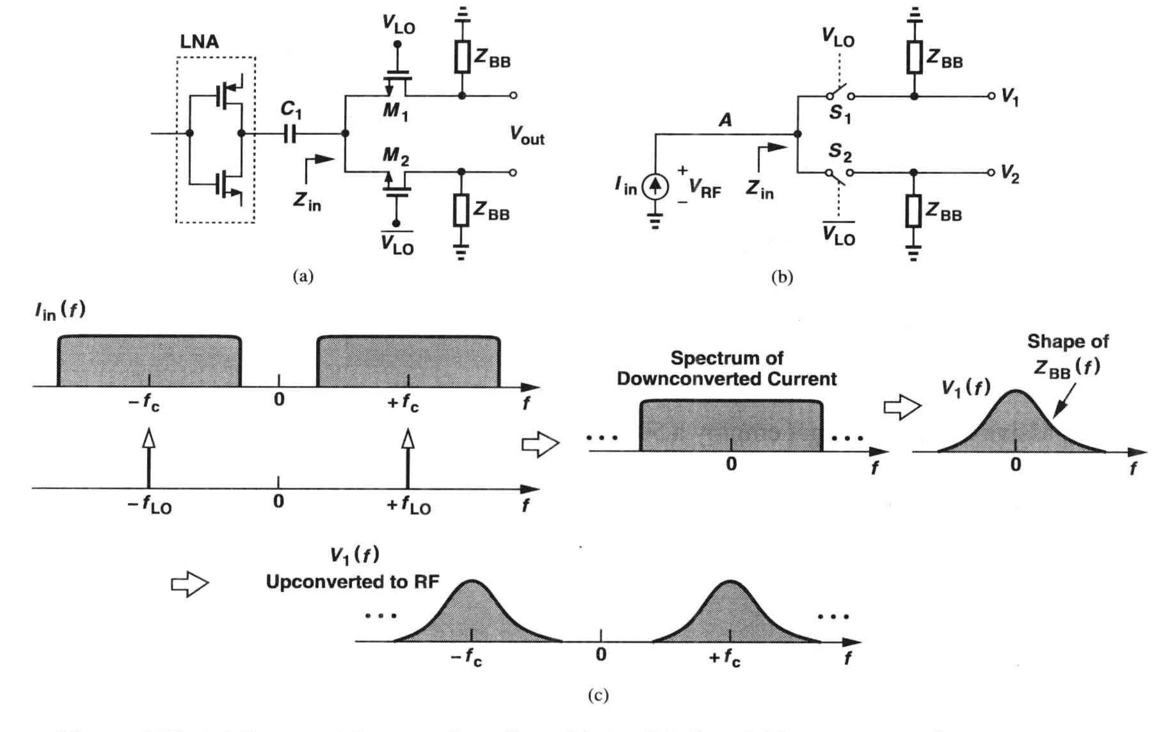


Figure 6.39 *(a) Current-driven passive mixer, (b) simplified model for input impedance calculation, (c) spectra at input and output.*

where $S(t)$ denotes a square wave toggling between 0 and 1, and $h(t)$ is the impulse response of Z_{BB}. In the frequency domain,

$$V_1(f) = [I_{in}(f) * S(f)] \cdot Z_{BB}(f), \tag{6.53}$$

where $S(f)$ is the spectrum of a square wave. As expected, upon convolution with the first harmonic of $S(f)$, $I_{in}(f)$ is translated to the baseband and is then subjected to the frequency response of $Z_{BB}(f)$. A similar phenomenon occurs in the lower arm.

We now make a critical observation [5]: the switches in Fig. 6.39(b) also mix the *baseband* waveforms with the LO, delivering the *upconverted* voltages to node A. Thus, $V_1(t)$ is multiplied by $S(t)$ as it returns to the input, and its spectrum is translated to RF. The spectrum of $V_2(t)$ is also upconverted and added to this result.

Figure 6.39(c) summarizes our findings, revealing that the downconverted spectrum of $I_{in}(f)$ is shaped by the frequency response of Z_{BB}, and the result "goes back" through the mixer, landing around f_c while retaining its spectral shape. In other words, in response to the spectrum shown for $I_{in}(f)$, an RF voltage spectrum has appeared at the input that is shaped by the baseband impedance. This implies that the input impedance around f_c resembles a frequency-translated version of $Z_{BB}(f)$. For example, if $Z_{BB}(f)$ is a low-pass impedance, then $Z_{in}(f)$ has a band-pass behavior [5].

The second property of current-driven passive mixers is that their noise and nonlinearity contribution are reduced [6]. This is because, ideally, a device in series with a *current source* does not alter the current passing through it.

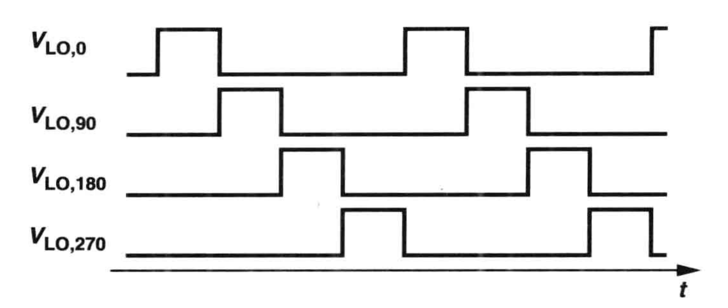


Figure 6.40 *Quadrature LO waveforms with 25% duty cycle.*

Passive mixers need not employ a 50% LO duty cycle. In fact, both voltage-driven and current-driven mixers utilizing a 25% duty cycle provide a higher gain. Figure 6.40 shows the quadrature LO waveforms according to this scenario. Writing the Fourier series for LO waveforms having a duty cycle of d, the reader can show that the RF current entering each switch generates an IF current given by [6]:

$$I_{IF}(t) = \frac{2}{\pi}\frac{\sin \pi d}{2d} I_{RF0} \cos \omega_{IF} t, \tag{6.54}$$

where I_{RF0} denotes the peak amplitude of the RF current. As expected, $d = 0.5$ yields a gain of $2/\pi$. More importantly, for $d = 0.25$, the gain reaches $2\sqrt{2}/\pi$, 3 dB higher. Of course, the generation of these waveforms becomes difficult at very high frequencies. [Ideally, we would choose $d \approx 0$ (impulse sampling) to raise this gain to unity.]

Another useful attribute of the 25% duty cycle in Fig. 6.40 is that the mixer switches driven by LO_0 and LO_{180} (or by LO_{90} and LO_{270}) are not on simultaneously. As a result, the mixer contributes smaller noise and nonlinearity [6].

6.3 ACTIVE DOWNCONVERSION MIXERS

Mixers can be realized so as to achieve conversion gain in *one* stage. Called active mixers, such topologies perform three functions: they convert the RF voltage to a current, "commutate" (steer) the RF current by the LO, and convert the IF current to voltage. These operations are illustrated in Fig. 6.41. While both passive and active mixers incorporate switching for frequency translation, the latter precede and follow the switching by voltage-to-current (V/I) and current-to-voltage (I/V) conversion, respectively, thereby achieving gain. We can intuitively observe that the input transconductance, I_{RF}/V_{RF}, and the output transresistance, V_{IF}/I_{IF}, can, in principle, assume arbitrarily large values, yielding an arbitrarily high gain.

Figure 6.42 depicts a typical single-balanced realization. Here, M_1 converts the input RF voltage to a current (and is hence called a "transconductor"), the differential pair M_2–M_3 commutates (steers) this current to the left and to the right, and R_1 and R_2 convert the output currents to voltage. We call M_2 and M_3 the "switching pair." As with our passive mixer study in Section 6.2, we wish to quantify the gain, noise, and nonlinearity of this circuit.

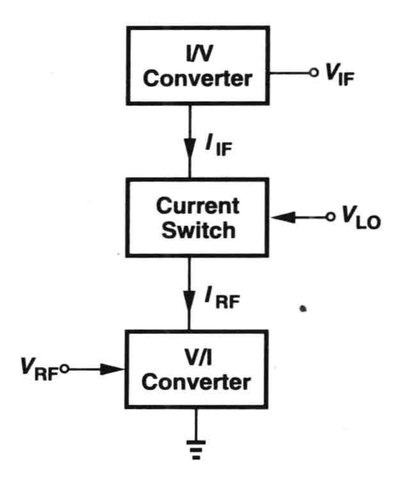


Figure 6.41 *Active mixer viewed as a V/I converter, a current switch, and an I/V converter.*

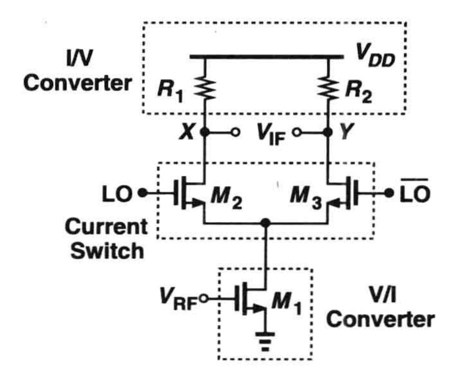


Figure 6.42 *Single-balanced active mixer.*

Note that the switching pair does not need rail-to-rail LO swings. In fact, as explained later, such swings degrade the linearity.

Double-Balanced Topology If the RF input is available in differential form, e.g., if the LNA provides differential outputs, then the active mixer of Fig. 6.42 must be modified accordingly. We begin by duplicating the circuit as shown in Fig. 6.43(a), where V_{RF}^+ and V_{RF}^- denote the differential phases of the RF input. Each half circuit commutates the RF current to its IF outputs. Since $V_{RF}^+ = -V_{RF}^-$, the small-signal IF components at X_1 and Y_1 are equal to the negative of those at X_2 and Y_2, respectively. That is, $V_{X1} = -V_{Y1} = -V_{X2} = V_{Y2}$, allowing us to short X_1 to Y_2 and X_2 to Y_1 and arrive at the double-balanced mixer in Fig. 6.43(b), where the load resistors are equal to $R_D/2$. We often draw the circuit as shown in Fig. 6.43(c) for the sake of compactness. Transistors M_2, M_3, M_5, and M_6 are called the "switching quad." We will study the advantages and disadvantages of this topology in subsequent sections.

One advantage of double-balanced mixers over their single-balanced counterparts stems from their rejection of amplitude noise in the LO waveform. We return to this property in Section 6.3.2.

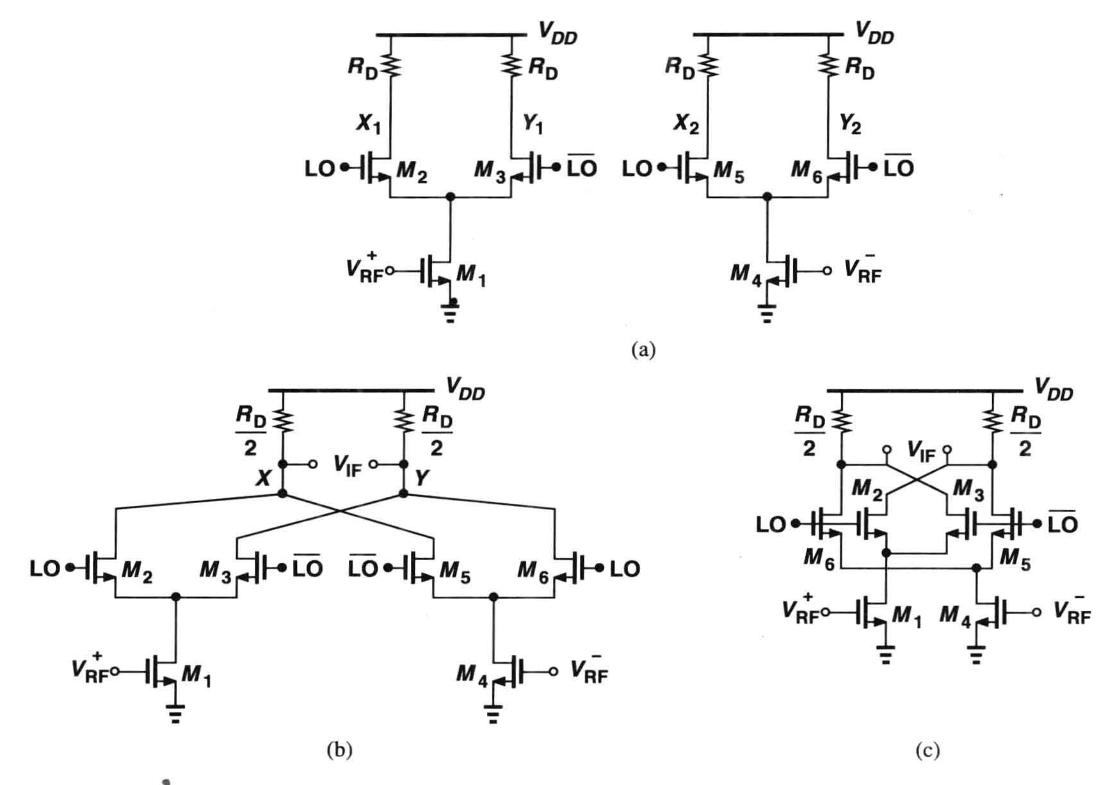


Figure 6.43 *(a) Two single-balanced mixers sensing differential RF inputs, (b) summation of output currents, (c) compact drawing of circuit.*

Example 6.11

Can the load resistors in the circuit of Fig. 6.43(b) be equal to R_D so as to double the gain?

Solution:

No, they cannot. Since the total bias current flowing through each resistor is doubled, R_D must be halved to comply with the voltage headroom.

6.3.1 Conversion Gain

In the circuit of Fig. 6.42, transistor M_1 produces a small-signal drain current equal to $g_{m1}V_{RF}$. With abrupt LO switching, the circuit reduces to that shown in Fig. 6.44(a), where M_2 multiplies I_{RF} by a square wave toggling between 0 and 1, $S(t)$, and M_3 multiplies I_{RF} by $S(t - T_{LO}/2)$ because LO and $\overline{\text{LO}}$ are complementary. It follows that

$$I_1 = I_{RF} \cdot S(t) \tag{6.55}$$

$$I_2 = I_{RF} \cdot S\left(t - \frac{T_{LO}}{2}\right). \tag{6.56}$$

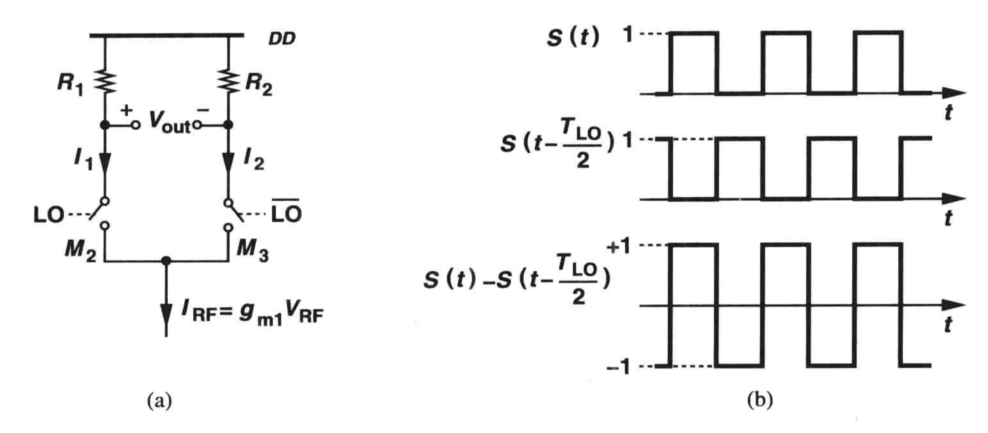


Figure 6.44 *(a) Equivalent circuit of active mixer, (b) switching waveforms.*

Since $V_{out} = V_{DD} - I_1R_1 - (V_{DD} - I_2R_2)$, we have for $R_1 = R_2 = R_D$,

$$V_{out}(t) = I_{RF}R_D\left[S\left(t - \frac{T_{LO}}{2}\right) - S(t)\right]. \tag{6.57}$$

From Fig. 6.44(b), we recognize that the switching operation in Eq. (6.57) is equivalent to multiplying I_{RF} by a square wave toggling between -1 and $+1$. Such a waveform exhibits a fundamental amplitude equal to $4/\pi$,[4] yielding an output given by

$$V_{out}(t) = I_{RF}(t)R_D \cdot \frac{4}{\pi}\cos\omega_{LO}t + \cdots. \tag{6.58}$$

If $I_{RF}(t) = g_{m1}V_{RF}\cos\omega_{RF}t$, then the IF component at $\omega_{RF} - \omega_{LO}$ is equal to

$$V_{IF}(t) = \frac{2}{\pi}g_{m1}R_DV_{RF}\cos(\omega_{RF} - \omega_{LO})t. \tag{6.59}$$

The voltage conversion gain is therefore equal to

$$\frac{V_{IF,p}}{V_{RF,p}} = \frac{2}{\pi}g_{m1}R_D. \tag{6.60}$$

What limits the conversion gain? We assume a given power budget, i.e., a certain bias current, I_{D1}, and show that the gain trades with the linearity and voltage headroom. The input transistor is sized according to the overdrive voltage, $V_{GS1} - V_{TH1}$, that yields the required IP_3 (Chapter 5). Thus, $V_{DS1,min} = V_{GS1} - V_{TH1}$. The transconductance of M_1 is limited by the current budget and IP_3, as expressed by $g_{m1} = 2I_{D1}/(V_{GS1} - V_{TH1})$ [or $I_{D1}/(V_{GS1} - V_{TH1})$ for velocity-saturated devices]. Also, the value of R_D is limited by the maximum allowable dc voltage across it. In other words, we must compute the minimum allowable value of V_X and V_Y in Fig. 6.42. As explained in Section 6.3.3, linearity

4. It is helpful to remember that the peak amplitude of the first harmonic of a square wave is *greater* than the peak amplitude of the square wave.

requirements dictate that M_2 and M_3 not enter the triode region so long as both carry current.

Suppose the gate voltages of M_2 and M_3 in Fig. 6.42 are held at the common-mode level of the differential LO waveforms, $V_{CM,LO}$ [Fig. 6.45(a)]. If M_1 is at the edge of saturation, then $V_N \geq V_{GS1} - V_{TH1}$:

$$V_{CM,LO} - V_{GS2,3} \geq V_{GS1} - V_{TH1}. \tag{6.61}$$

Now consider the time instant at which the gate voltages of M_2 and M_3 reach $V_{CM,LO} + V_0$ and $V_{CM,LO} - V_0$, respectively, where $V_0 = \sqrt{2}(V_{GS2,3} - V_{TH2})/2$, a value high enough to turn off M_3 [Fig. 6.45(b)]. For M_2 to remain in saturation up to this point, its drain voltage must not fall below $V_{CM,LO} + \sqrt{2}(V_{GS2,3} - V_{TH2})/2 - V_{TH2}$:

$$V_{X,min} = V_{CM,LO} + \frac{\sqrt{2}}{2}(V_{GS2,3} - V_{TH2}) - V_{TH2}, \tag{6.62}$$

which, from Eq. (6.61), reduces to

$$V_{X,min} = V_{GS1} - V_{TH1} + \left(1 + \frac{\sqrt{2}}{2}\right)(V_{GS2,3} - V_{TH2}). \tag{6.63}$$

Thus, $V_{X,min}$ must accommodate the overdrive of M_1 and about 1.7 times the "equilibrium" overdrive of each of the switching transistors. The maximum allowable dc voltage across each load resistor is equal to

$$V_{R,max} = V_{DD} - \left[V_{GS1} - V_{TH1} + \left(1 + \frac{\sqrt{2}}{2}\right)(V_{GS2,3} - V_{TH2})\right]. \tag{6.64}$$

Since each resistor carries half of I_{D1},

$$R_{D,max} = \frac{2V_{R,max}}{I_{D1}}. \tag{6.65}$$

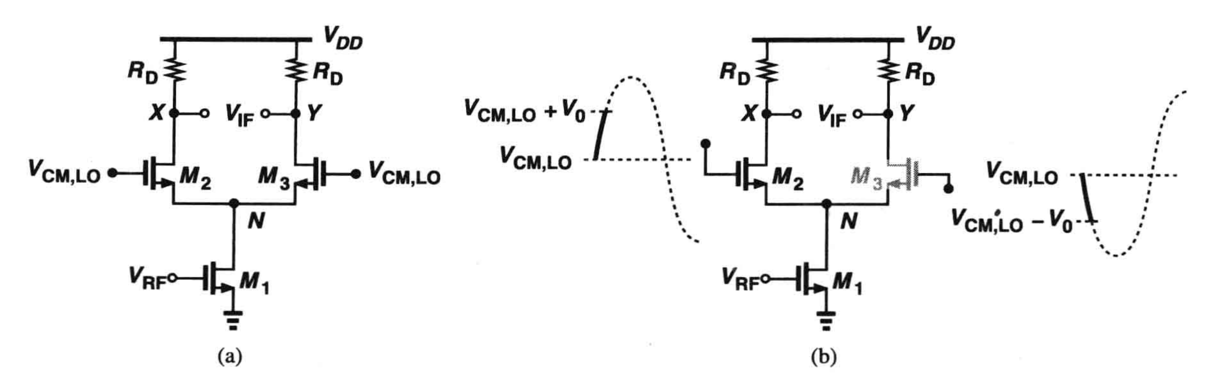


Figure 6.45 *(a) Active mixer with LO at CM level, (b) required swing to turn one device off.*

From (6.64) and (6.65), we obtain the maximum voltage conversion gain as

$$A_{V,max} = \frac{2}{\pi} g_{m1} R_{D,max} \tag{6.66}$$

$$= \frac{8}{\pi} \frac{V_{R,max}}{V_{GS1} - V_{TH1}}. \tag{6.67}$$

We therefore conclude that low supply voltages severely limit the gain of active mixers.

Example 6.12

A single-balanced active mixer requires an overdrive voltage of 300 mV for the input V/I converter transistor. If each switching transistor has an equilibrium overdrive of 150 mV and the peak LO swing is 300 mV, how much conversion gain can be obtained with a 1-V supply?

Solution:

From Eq. (6.64), $V_{R,max} = 444$ mV and hence

$$A_{V,max} = 3.77 \tag{6.68}$$

$$\approx 11.5\,\text{dB}. \tag{6.69}$$

Owing to the relatively low conversion gain, the noise contributed by the load resistors and following stages may become significant.

How much room for improvement do we have? Given by IP_3 requirements, the overdrive of the input transistor has little flexibility unless the gain of the preceding LNA can be reduced. This is possible if the mixer noise figure can also be lowered, which, as explained in Section 6.3.2, trades with the power dissipation and input capacitance of the mixer. The equilibrium overdrive of the switching transistors can be reduced by making the two transistors wider (while raising the capacitance seen at the LO port).

The conversion gain may also fall if the LO swing is lowered. As illustrated in Fig. 6.46, while M_2 and M_3 are near equilibrium, the RF current produced by M_1 is

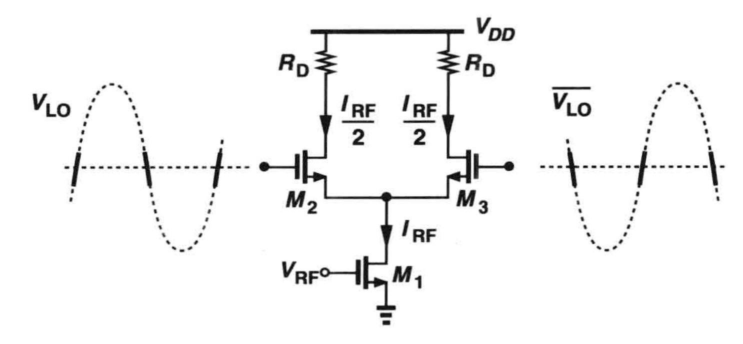


Figure 6.46 *RF current as a CM component near LO zero crossings.*

split approximately equally between them, thus appearing as a *common-mode* current and yielding little conversion gain for that period of time. Reduction of the LO swing tends to increase this time and lower the gain (unless the LO is a square wave).

Example 6.13

Figure 6.47 shows a "dual-gate mixer," where M_1 and M_2 can be viewed as one transistor with two gates. Identify the drawbacks of this circuit.

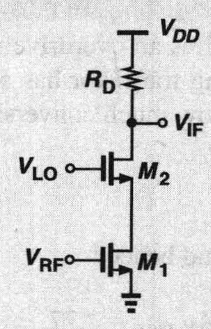


Figure 6.47 *Dual-gate mixer.*

Solution:

For M_2 to operate as a switch, its gate voltage must fall to V_{TH2} above zero (why?) regardless of the overdrive voltages of the two transistors. For this reason, the dual-gate mixer typically calls for larger LO swings than the single-balanced active topology does. Furthermore, since the RF current of M_1 is now multiplied by a square wave toggling between 0 and 1, the conversion gain is half:

$$A_V = \frac{1}{\pi} g_{m1} R_D. \tag{6.70}$$

Additionally, all of the frequency components produced by M_1 appear at the output without translation because they are multiplied by the average value of the square wave, $1/2$. Thus, half of the flicker noise of M_1—a high-frequency device and hence small—emerges at IF. Also, low-frequency beat components resulting from even-order distortion (Chapter 4) in M_1 directly corrupt the output, leading to a low IP_2. The dual-gate mixer does not require differential LO waveforms, a minor advantage. For these reasons, this topology is rarely used in modern RF design.

With a sinusoidal LO, the drain currents of the switching devices depart from square waves, remaining approximately equal for a fraction of each half cycle, ΔT [Fig. 6.48(a)]. As mentioned previously, the circuit exhibits little conversion gain during these periods. We now wish to estimate the reduction in the gain.

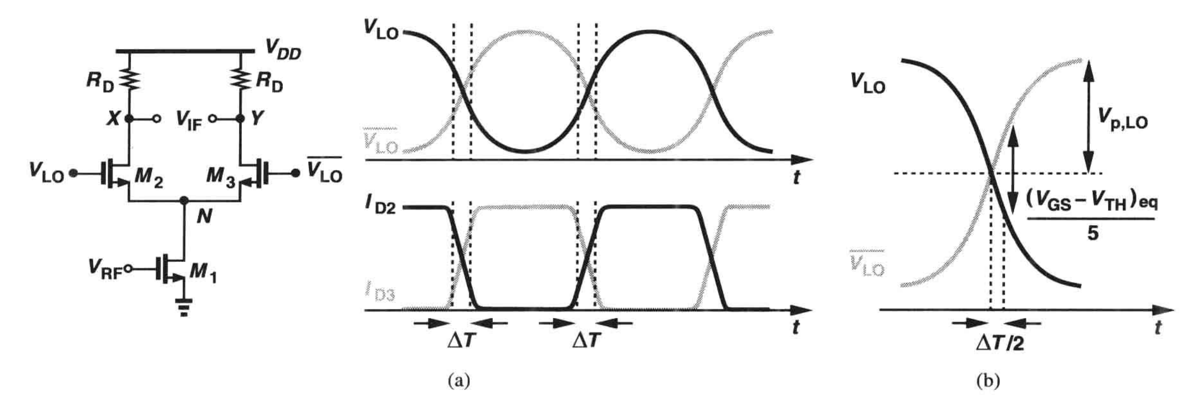


Figure 6.48 *(a) Effect of gradual LO transitions, (b) magnified LO waveforms.*

A differential pair having an equilibrium overdrive of $(V_{GS} - V_{TH})_{eq}$ steers most of its tail current for a differential input voltage, ΔV_{in}, of $\sqrt{2}(V_{GS} - V_{TH})_{eq}$ (for square-law devices). We assume that the drain currents are roughly equal for $\Delta V_{in} \leq (V_{GS} - V_{TH})_{eq}/5$ and calculate the corresponding value of ΔT. We note from Fig. 6.48(b) that, if each single-ended LO waveform has a peak amplitude of $V_{p,LO}$, then LO and $\overline{\text{LO}}$ reach a difference of $(V_{GS} - V_{TH})_{eq}/5$ in approximately $\Delta T/2 = (V_{GS} - V_{TH})_{eq}/5/(2V_{p,LO}\omega_{LO})$ seconds. Multiplying this result by a factor of 4 to account for the total time on both rising and falling edges and normalizing to the LO period, we surmise that the overall gain of the mixer is reduced to

$$A_V = \frac{2}{\pi} g_{m1} R_D \left(1 - \frac{2\Delta T}{T_{LO}}\right) \tag{6.71}$$

$$= \frac{2}{\pi} g_{m1} R_D \left[1 - \frac{(V_{GS} - V_{TH})_{eq}}{5\pi V_{p,LO}}\right]. \tag{6.72}$$

Example 6.14

Repeat Example 6.12 but take the gradual LO edges into account.

Solution:

The gain expressed by Eq. (6.68) must be multiplied by $1 - 0.0318 \approx 0.97$:

$$A_{V,max} \approx 3.66 \tag{6.73}$$

$$\approx 11.3\,\text{dB}. \tag{6.74}$$

Thus, the gradual LO transitions lower the gain by about 0.2 dB.

The second phenomenon that degrades the gain relates to the total capacitance seen at the drain of the input transistor. Consider an active mixer in one-half of the LO cycle (Fig. 6.49). With abrupt LO edges, M_2 is on and M_3 is off, yielding a total capacitance at

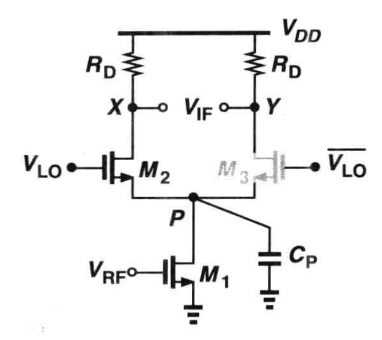


Figure 6.49 *Loss of RF current to ground through C_P.*

node P equal to

$$C_P = C_{DB1} + C_{GS2} + C_{GS3} + C_{SB2} + C_{SB3}. \tag{6.75}$$

Note that C_{GS3} is substantially smaller than C_{GS2} in this phase (why?). The RF current produced by M_1 is split between C_P and the resistance seen at the source of M_2, $1/g_{m2}$ (if body effect is neglected). Thus, the voltage conversion gain is reduced by a factor of $g_{m2}/(sC_P + g_{m2})$; i.e., Eq. (6.72) must be modified as

$$A_{V,max} = \frac{2}{\pi} g_{m1} R_D \left[1 - \frac{2(V_{GS} - V_{TH})_{eq}}{5\pi V_{P,LO}}\right] \frac{g_{m2}}{\sqrt{C_P^2\omega^2 + g_{m2}^2}}. \tag{6.76}$$

How significant is this current division? In other words, how does $C_P^2\omega^2$ compare with g_{m2}^2 in the above expression? Note that g_{m2}/C_P is well below the maximum f_T of M_2 because (a) the sum of C_{DB1}, C_{SB2}, C_{SB3}, and C_{GS3} is comparable with or larger than C_{GS2}, and (b) the low overdrive voltage of M_2 (imposed by headroom and gain requirements) also leads to a low f_T. We therefore observe that the effect of C_P may become critical for frequencies higher than roughly one-tenth of the maximum f_T of the transistors.

Example 6.15

If the output resistance of M_2 in Fig. 6.49 is not neglected, how should it be included in the calculations?

Solution:

Since the output frequency of the mixer is much lower than the input and LO frequencies, a capacitor is usually tied from each output node to ground to filter the unwanted components (Fig. 6.50). As a result, the resistance seen at the source of M_2 in Fig. 6.50 is simply equal to $(1/g_{m2})||r_{O2}$ because the output capacitor establishes an ac ground at the drain of M_2 at the input frequency.

Example 6.15 *(Continued)*

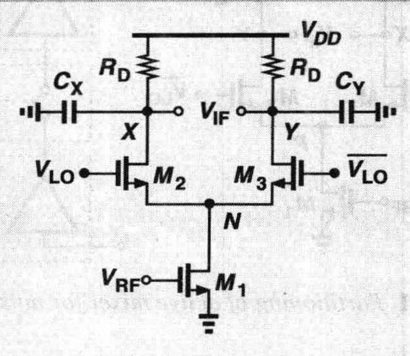


Figure 6.50 *Capacitors tied to output nodes to limit the bandwidth.*

Example 6.16

Compare the voltage conversion gains of single-balanced and double-balanced active mixers.

Solution:

From Fig. 6.43(a), we recognize that $(V_{X1} - V_{Y1})/V_{RF}^+$ is equal to the voltage conversion gain of a single-balanced mixer. Also, $V_{X1} = V_{Y2}$ and $V_{Y1} = V_{X2}$ if $V_{RF}^- = -V_{RF}^+$. Thus, if Y_2 is shorted to X_1, and X_2 to Y_1, these node voltages remain unchanged. In other words, $V_X - V_Y$ in Fig. 6.43(b) is equal to $V_{X1} - V_{Y1}$ in Fig. 6.43(a). The differential voltage conversion gain of the double-balanced topology is therefore given by

$$\frac{V_X - V_Y}{V_{RF}^+ - V_{RF}^-} = \frac{V_{X1} - V_{Y1}}{2V_{RF}^+}, \tag{6.77}$$

which is half of that of the single-balanced counterpart. This reduction arises because the limited voltage headroom disallows a load resistance of R_D in Fig. 6.43(b) (Example 6.11).

6.3.2 Noise in Active Mixers

The analysis of noise in active mixers is somewhat different from the study undertaken in Section 6.2.3 for passive mixers. As illustrated conceptually in Fig. 6.51, the noise components of interest lie in the RF range before downconversion and in the IF range after downconversion. Note that the frequency translation of RF noise by the switching devices prohibits the direct use of small-signal ac and noise analysis in circuit simulators (as is done for LNAs), necessitating simulations in the time domain. Moreover, the noise contributed by the switching devices exhibits time-varying statistics, complicating the analysis.

Qualitative Analysis To gain insight into the noise behavior of active mixers, we begin with a qualitative study. Let us first assume abrupt LO transitions and consider the

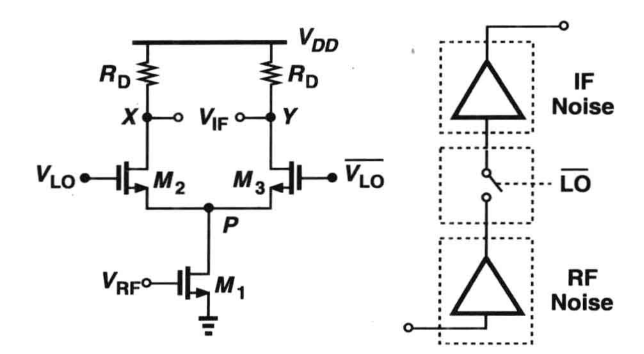


Figure 6.51 *Partitioning of active mixer for noise analysis.*

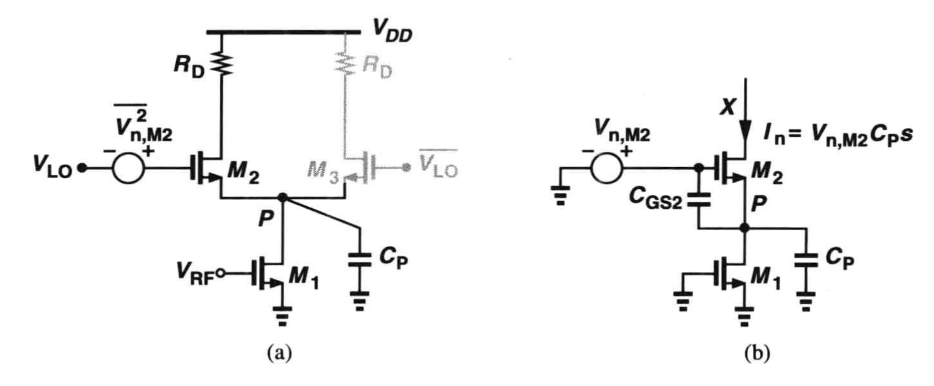


Figure 6.52 *(a) Effect of noise when one transistor is off, (b) equivalent circuit of (a).*

representation in Fig. 6.52(a) for half of the LO cycle. Here,

$$C_P = C_{GD1} + C_{DB1} + C_{SB2} + C_{SB3} + C_{GS3}. \tag{6.78}$$

In this phase, the circuit reduces to a cascode structure, with M_2 contributing some noise because of the capacitance at node P (Chapter 5). Recall from the analysis of cascode LNAs in Chapter 5 that, at frequencies well below f_T, the output noise current generated by M_2 is equal to $V_{n,M2}C_P s$ [Fig. 6.52(b)]. This noise and the noise current of M_1 (which is dominant) are multiplied by a square wave toggling between 0 and 1. Transistor M_3 plays an identical role in the next half cycle of the LO.

Now consider a more realistic case where the LO transitions are not abrupt, allowing M_2 and M_3 to remain on simultaneously for part of the period. As depicted in Fig. 6.53, the circuit now resembles a differential pair near equilibrium, amplifying the noise of M_2 and M_3—while the noise of M_1 has little effect on the output because it behaves as a common-mode disturbance.

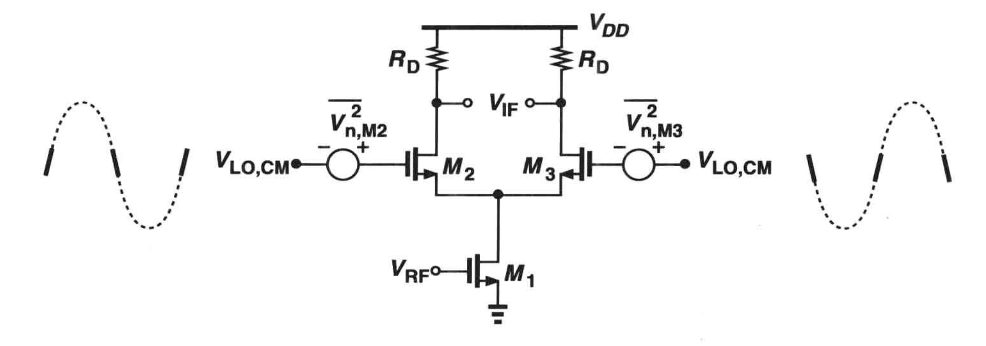


Figure 6.53 *Effect of noise of M_2 and M_3 near equilibrium.*

Example 6.17

Compare single-balanced and double-balanced active mixers in terms of their noise behavior. Assume the latter's total bias current is twice the former's.

Solution:

Let us first study the output noise *currents* of the mixers [Fig. 6.54(a)]. If the total differential output noise current of the single-balanced topology is $\overline{I^2_{n,sing}}$, then that of the double-balanced circuit is equal to $\overline{I^2_{n,doub}} = 2\overline{I^2_{n,sing}}$ (why?). Next, we determine the output noise voltages, bearing in mind that the load resistors differ by a factor of two [Fig. 6.54(b)]. We have

$$\overline{V^2_{n,out,sing}} = \overline{I^2_{n,sing}}(R_D)^2 \tag{6.79}$$

$$\overline{V^2_{n,out,doub}} = \overline{I^2_{n,doub}}\left(\frac{R_D}{2}\right)^2. \tag{6.80}$$

But recall from Example 6.16 that the voltage conversion gain of the double-balanced mixer is half of that of the single-balanced topology. Thus, the input-referred noise voltages of the two circuits are related by

$$\overline{V^2_{n,in,sing}} = \frac{1}{2}\overline{V^2_{n,in,doub}}. \tag{6.81}$$

In this derivation, we have not included the noise of the load resistors. The reader can show that Eq. (6.81) remains valid even with their noise taken into account. The single-balanced mixer therefore exhibits less input noise and consumes less power.

(*Continues*)

Example 6.17 *(Continued)*

Figure 6.54 *(a) Output noise currents of single-balanced and double-balanced mixers, (b) corresponding output noise voltages.*

It is important to make an observation regarding the mixer of Fig. 6.53. The noise generated by the local oscillator and its buffer becomes indistinguishable from the noise of M_2 and M_3 when these two transistors are around equilibrium. As depicted in Fig. 6.55, a differential pair serving as the LO buffer may produce an output noise much higher than

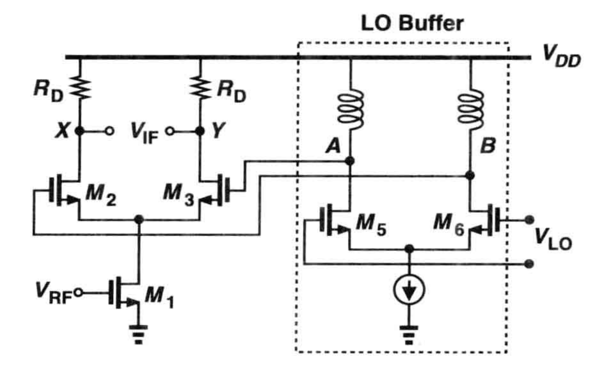


Figure 6.55 *Effect of LO buffer noise on single-balanced mixer.*

that of M_2 and M_3. It is therefore necessary to simulate the noise behavior of mixers with the LO circuitry present.

Example 6.18

Study the effect of LO noise on the performance of double-balanced active mixers.

Solution:

Drawing the circuit as shown in Fig. 6.56, we note that the LO noise voltage is converted to current by each switching pair and summed with *opposite polarities*. Thus, the double-balanced topology is much more immune to LO noise—a useful property obtained at the cost of the 3-dB noise penalty expressed by Eq. (6.81) and the higher power dissipation. Here, we have assumed that the noise components in LO and $\overline{\text{LO}}$ are *differential*. We study this point in Problem 6.6, concluding that this assumption is reasonable for a true differential buffer but not for a quasi-differential circuit.

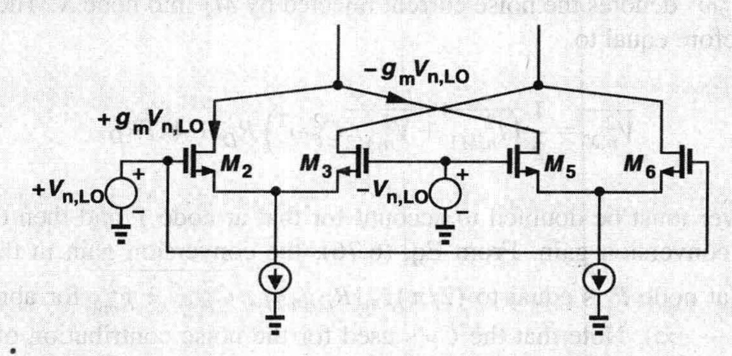


Figure 6.56 *Effect of LO noise on double-balanced mixer.*

Quantitative Analysis Consider the single-balanced mixer depicted in Fig. 6.51. From our qualitative analysis, we identify three sections in the circuit: the RF section, the time-varying (switching) section, and the IF section. To estimate the input-referred noise voltage, we apply the following procedure: (1) for each source of noise, determine a "conversion gain" to the IF output; (2) multiply the magnitude of each noise by the corresponding gain and add up all of the resulting powers, thus obtaining the total noise at the IF output; (3) divide the output noise by the overall conversion gain of the mixer to refer it to the input.

Let us begin the analysis by assuming abrupt LO transitions with a 50% duty cycle. In each half cycle of the LO, the circuit resembles that in Fig. 6.57, i.e., the noise of M_1 ($I_{n1,M1}$) and each of the switching devices is multiplied by a square wave toggling between 0 and 1. We have seen in Example 6.4 that, if white noise is switched on and off with 50% duty cycle, the resulting spectrum is still white while carrying half of the power. Thus, half of the noise powers (squared current quantities) of M_1 and M_2 is injected into

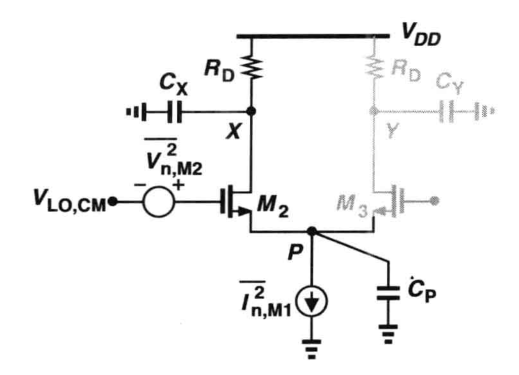


Figure 6.57 *Noise of input device and one switching device in an active mixer.*

node X, generating an output noise spectral density given by $(1/2)(\overline{I^2_{n,M1}} + \overline{V^2_{n,M2}}C_P^2\omega^2)R_D^2$, where $\overline{V^2_{n,M2}}C_P^2\omega^2$ denotes the noise current injected by M_2 into node X. The total noise at node X is therefore equal to

$$\overline{V^2_{n,X}} = \frac{1}{2}\left(\overline{I^2_{n,M1}} + \overline{V^2_{n,M2}}C_P^2\omega^2\right)R_D^2 + 4kTR_D. \tag{6.82}$$

The noise power must be doubled to account for that at node Y and then divided by the square of the conversion gain. From Eq. (6.76), the conversion gain in the presence of a capacitance at node P is equal to $(2/\pi)g_{m1}R_D g_{m2}/\sqrt{C_P^2\omega^2 + g_{m2}^2}$ for abrupt LO edges (i.e., if $V_{p,LO} \to \infty$). Note that the C_P's used for the noise contribution of M_2 and gain calculation are given by (6.75) and (6.78), respectively, and slightly different. Nonetheless, we assume they are approximately equal. The input-referred noise voltage is therefore given by

$$\overline{V^2_{n,in}} = \frac{\left(4kT\gamma g_{m1} + \dfrac{4kT\gamma}{g_{m2}}C_P^2\omega^2\right)R_D^2 + 8kTR_D}{\dfrac{4}{\pi^2}g_{m1}^2R_D^2\dfrac{g_{m2}^2}{C_P^2\omega^2 + g_{m2}^2}} \tag{6.83}$$

$$= \pi^2\left(\frac{C_P^2\omega^2}{g_{m2}^2} + 1\right)kT\left(\frac{\gamma}{g_{m1}} + \frac{\gamma C_P^2\omega^2}{g_{m2}g_{m1}^2} + \frac{2}{g_{m1}^2R_D}\right). \tag{6.84}$$

If the effect of C_P is negligible, then

$$\overline{V^2_{n,in}} = \pi^2 kT\left(\frac{\gamma}{g_{m1}} + \frac{2}{g_{m1}^2R_D}\right). \tag{6.85}$$

Example 6.19

Compare Eq. (6.85) with the input-referred noise voltage of a common-source stage having the same transconductance and load resistance.

Solution:

For the CS stage, we have

$$\overline{V_{n,in,CS}^2} = 4kT\left(\frac{\gamma}{g_{m1}} + \frac{1}{g_{m1}^2 R_D}\right). \quad (6.86)$$

Thus, even if the second term in the parentheses is negligible, the mixer exhibits 3.92 dB higher noise power. With a finite C_P and LO transition times, this difference becomes even larger.

The term $\pi^2 kT\gamma/g_{m1}$ in (6.85) represents the input-referred contribution of M_1. This appears puzzling: why is this contribution simply not equal to the gate-referred noise of M_1, $4kT\gamma/g_{m1}$? We investigate this point in Problem 6.7.

We now consider the effect of gradual LO transitions on the noise behavior. Similar to the gain calculations in Section 6.3.1, we employ a piecewise-linear approximation (Fig. 6.58): the switching transistors are considered near equilibrium for $2\Delta T = 2(V_{GS} - V_{TH})_{eq}/(5V_{P,LO}\omega_{LO})$ seconds per LO cycle, injecting noise to the output as a differential pair. During this time period, M_1 contributes mostly common-mode noise, and the output noise is equal to

$$\overline{V_{n,diff}^2} = 2(4kT\gamma g_{m2}R_D^2 + 4kTR_D), \quad (6.87)$$

where we assume $g_{m2} \approx g_{m3}$. Now, this noise power must be weighted by a factor of $2\Delta T/T_{LO}$, and that in the numerator of Eq. (6.83) by a factor of $1 - 2\Delta T/T_{LO}$. The sum of

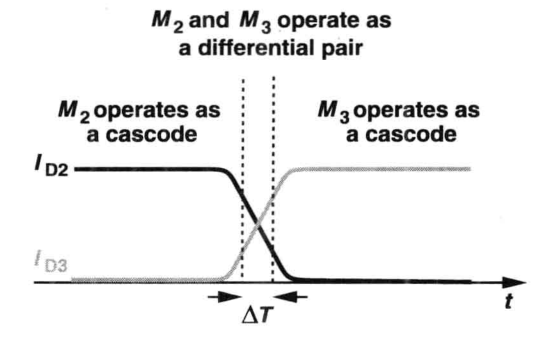


Figure 6.58 *Piecewise-linear waveforms for mixer noise calculation.*

these weighted noise powers must then be divided by the square of (6.76) to refer it to the input. The input-referred noise is thus given by

$$\overline{V_{n,in}^2} = \frac{8kT(\gamma g_{m2}R_D^2 + R_D)\dfrac{2\Delta T}{T_{LO}} + \left[4kT\gamma\left(g_{m1} + \dfrac{C_P^2\omega^2}{g_{m2}}\right)R_D^2 + 8kTR_D\right]\left(1 - \dfrac{2\Delta T}{T_{LO}}\right)}{\dfrac{4}{\pi^2}g_{m1}^2 R_D^2 \dfrac{g_{m2}^2}{C_P^2\omega^2 + g_{m2}^2}\left(1 - \dfrac{2\Delta T}{T_{LO}}\right)^2}. \tag{6.88}$$

Equation (6.88) reveals that the equilibrium overdrive voltage of the switching devices plays a complex role here: (1) in the first term in the numerator, $g_{m2} \propto (V_{GS} - V_{TH})_{eq}^{-1}$ (for a given bias current), whereas $\Delta T \propto (V_{GS} - V_{TH})_{eq}$; (2) the noise power expressed by the second term in the numerator is proportional to $1 - 2\Delta T/T_{LO}$ while the squared gain in the denominator varies in proportion to $(1 - 2\Delta T/T_{LO})^2$, suggesting that ΔT must be minimized.

Example 6.20

A single-balanced mixer is designed for a certain IP_3, bias current, LO swing, and supply voltage. Upon calculation of the noise, we find it unacceptably high. What can be done to lower the noise?

Solution:

The overdrive voltages and the dc drop across the load resistors offer little flexibility. We must therefore sacrifice power for noise by a direct scaling of the design. Illustrated in Fig. 6.59, the idea is to scale the transistor widths *and* currents by a factor of α and the load resistors by a factor of $1/\alpha$. All of the voltages thus remain unchanged, but the input-referred noise voltage, $\sqrt{\overline{V_{n,in}^2}}$, falls by a factor of $\sqrt{\alpha}$. Unfortunately, this scaling also scales the capacitances seen at the RF and LO ports, making the design of the LNA and the LO buffer more difficult and/or more power-hungry.

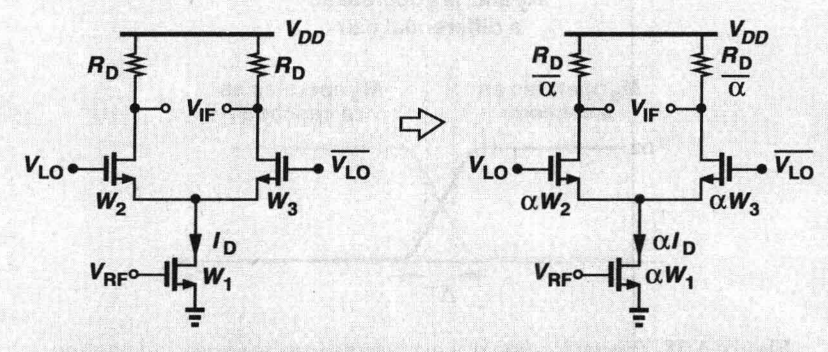


Figure 6.59 *Effect of scaling on noise.*

Flicker Noise Unlike passive mixers, active topologies suffer from substantial flicker noise at their output, a serious issue if the IF signal resides near zero frequency and has a narrow bandwidth.

Consider the circuit shown in Fig. 6.60(a). With perfect symmetry, the $1/f$ noise of I_{SS} does not appear at the output because it is mixed with ω_{LO} (and its harmonics). Thus, only the flicker noise of M_2 and M_3 must be considered. The noise of M_2, $\overline{V_{n2}^2}$, experiences the gain of the differential pair as it propagates to the output. Fortunately, the large LO swing heavily saturates (desensitizes) the differential pair most of the time, thereby lowering the gain seen by $\overline{V_{n2}^2}$.

In order to compute the gain experienced by V_{n2} in Fig. 6.60(a), we assume a sinusoidal LO but also a small switching time for M_2 and M_3 such that I_{SS} is steered almost instantaneously from one to the other at the zero crossings of LO and $\overline{\text{LO}}$ [Fig. 6.60(b)]. How does V_{n2} alter this behavior? Upon addition to the LO waveform, the noise modulates the zero crossings of the LO [7]. This can be seen by computing the time at which the gate voltages of M_1 and M_2 are equal; i.e., by equating the instantaneous gate voltages of M_2 and M_3:

$$V_{CM} + V_{p,LO} \sin \omega_{LO} t + V_{n2}(t) = V_{CM} - V_{p,LO} \sin \omega_{LO} t, \tag{6.89}$$

obtaining

$$2V_{p,LO} \sin \omega_{LO} t = -V_{n2}(t). \tag{6.90}$$

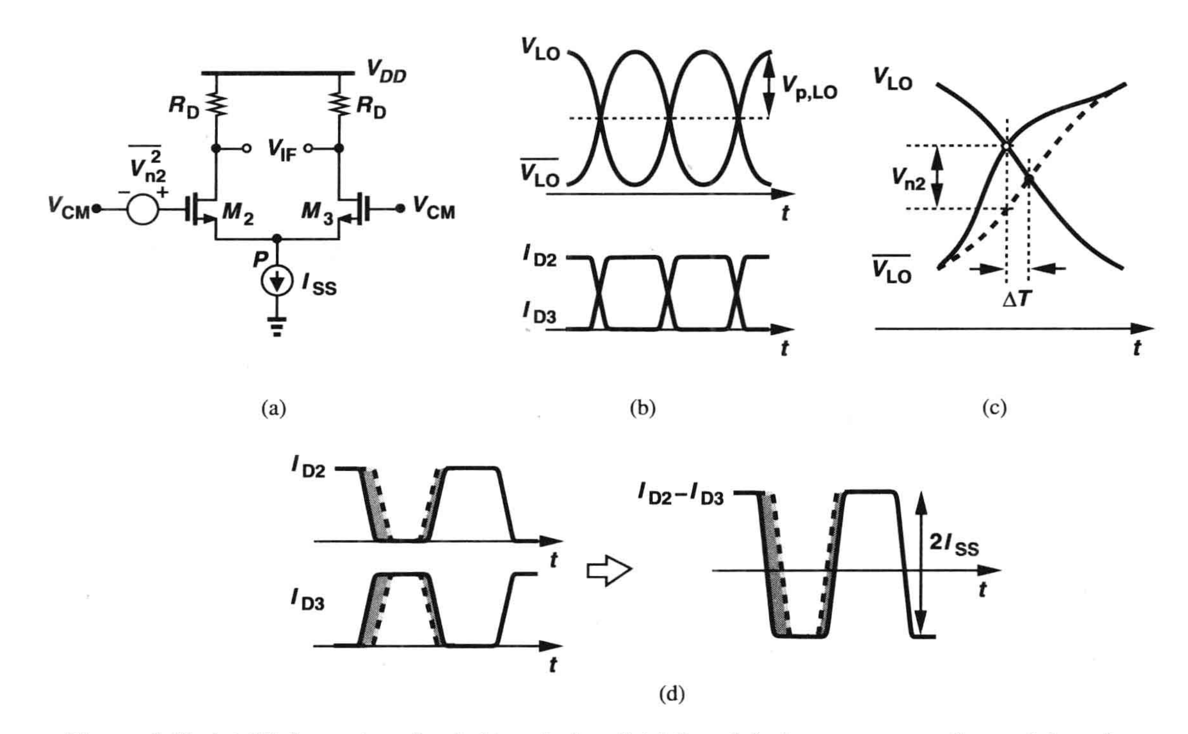


Figure 6.60 *(a) Flicker noise of switching device, (b) LO and drain current waveforms, (c) modulation of zero crossing due to flicker noise, (d) equivalent pulsewidth modulation.*

In the vicinity of $t = 0$, we have

$$2V_{p,LO}\omega_{LO}t \approx -V_{n2}(t). \tag{6.91}$$

The crossing of LO and $\overline{\text{LO}}$ is displaced from its ideal point by an amount of ΔT ($\omega_{LO}\Delta T \ll 1$ rad) [Fig. 6.60(c)]:

$$2V_{p,LO}\omega_{LO}\Delta T \approx -V_{n2}(t). \tag{6.92}$$

That is,

$$|\Delta T| = \frac{|V_{n2}(t)|}{2V_{p,LO}\omega_{LO}}. \tag{6.93}$$

Note that $2V_{p,LO}\omega_{LO}$ is the slope of the *differential* LO waveform,[5] S_{LO}, and hence $|\Delta T| = |V_{n2}(t)|/S_{LO}$.

We now assume nearly abrupt drain current switching for M_2 and M_3 and consider the above zero-crossing deviation as *pulsewidth* modulation of the currents [Fig. 6.60(d)]. Drawing the differential output current as in Fig. 6.60(d), we note that the modulated output is equal to the ideal output plus a noise waveform consisting of a series of narrow pulses of height $2I_{SS}$ and width ΔT and occurring *twice* per period [7]. If each narrow pulse is approximated by an impulse, the noise waveform in $I_{D2} - I_{D3}$ can be expressed as

$$I_{n,out}(t) = \sum_{k=-\infty}^{+\infty} \frac{2I_{SS}V_{n2}(t)}{S_{LO}}\delta\left(t - k\frac{T_{LO}}{2}\right). \tag{6.94}$$

In the frequency domain, from Eq. (6.9),

$$I_{n,out}(f) = \frac{4I_{SS}}{T_{LO}S_{LO}} \sum_{k=-\infty}^{+\infty} V_{n2}(f)\delta(t - 2kf_{LO}). \tag{6.95}$$

The baseband component is obtained for $k = 0$ because $V_{n2}(f)$ has a low-pass spectrum. It follows that

$$I_{n,out}(f)|_{k=0} = \frac{I_{SS}}{\pi V_{p,LO}}V_{n2}(f), \tag{6.96}$$

and hence

$$V_{n,out}(f)|_{k=0} = \frac{I_{SS}R_D}{\pi V_{p,LO}}V_{n2}(f). \tag{6.97}$$

In other words, the flicker noise of each transistor is scaled by a factor of $I_{SS}R_D/(\pi V_{p,LO})$ as it appears at the output. It is therefore desirable to *minimize* the bias current of the switching devices. Note that this quantity must be multiplied by $\sqrt{2}$ to account for the flicker noise of M_3 as well.

---

5. Because the *difference* between V_{LO} and $\overline{V_{LO}}$ must reach zero in ΔT seconds.

Example 6.21

Refer the noise found above to the input of the mixer.

Solution:

Multiplying the noise by $\sqrt{2}$ to account for the noise of M_3 and dividing by the conversion gain, $(2/\pi)g_{m1}R_D$, we have

$$V_{n,in}(f)|_{k=0} = \frac{\sqrt{2}I_{SS}}{2g_{m1}V_{p,LO}}V_{n2}(f) \tag{6.98}$$

$$= \frac{\sqrt{2}(V_{GS}-V_{TH})_1}{4V_{p,LO}}V_{n2}(f). \tag{6.99}$$

For example, if $(V_{GS}-V_{TH})_1 = 250\,\text{mV}$ and $V_{p,LO} = 300\,\text{mV}$, then $V_{n2}(f)$ is reduced by about a factor of 3.4 when referred to the input. Note, however, that (1) $V_{n2}(f)$ is typically very large because M_2 and M_3 are relatively small, and (2) the noise voltage found above must be multiplied by $\sqrt{2}$ to account for the noise of M_3.

The above study also explains the low $1/f$ noise of *passive* mixers. Since $I_{SS} = 0$ in passive topologies, a noise voltage source in series with the gate experiences a high attenuation as it appears at the output. (Additionally, MOSFETs carrying negligible current produce negligible flicker noise.)

The reader may wonder if the above results apply to the thermal noise of M_2 and M_3 as well. Indeed, the analysis is identical [7] and the same results are obtained, with $V_{n2}(f)$ replaced with $4kT\gamma/g_{m2}$. The reader can show that this method and our earlier method of thermal noise analysis yield roughly equal results if $\pi V_{p,LO} \approx 5(V_{GS}-V_{TH})_{eq2,3}$.

Another flicker noise mechanism in active mixers arises from the finite capacitance at node P in Fig. 6.60(a) [7]. It can be shown that the differential output current in this case includes a flicker noise component given by [7]

$$I_{n,out}(f) = 2f_{LO}C_PV_{n2}(f). \tag{6.100}$$

Thus, a higher tail capacitance or LO frequency intensifies this effect. Nonetheless, the first mechanism tends to dominate at low and moderate frequencies.

6.3.3 Linearity

The linearity of active mixers is determined primarily by the input transistor's overdrive voltage. As explained in Chapter 5, the IP_3 of a common-source transistor rises with the overdrive, eventually reaching a relatively constant value.

The input transistor imposes a direct trade-off between nonlinearity and noise because

$$IP_3 \propto V_{GS}-V_{TH} \tag{6.101}$$

$$\overline{V_{n,in}^2} = \frac{4kT\gamma}{g_m} = \frac{4kT\gamma}{2I_D}(V_{GS}-V_{TH}). \tag{6.102}$$

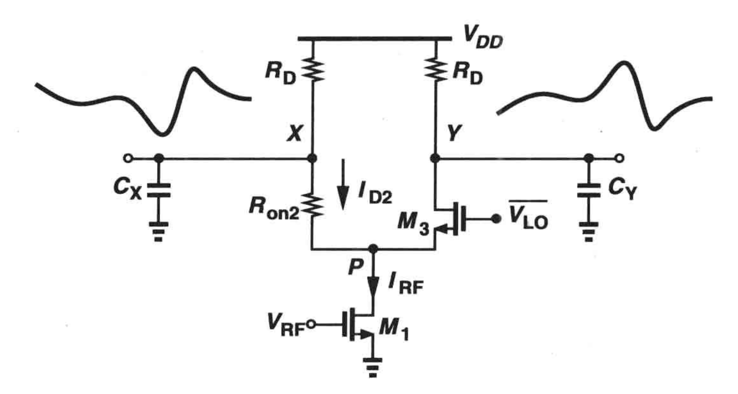


Figure 6.61 *Effect of output waveform on current steering when one device enters the triode region.*

We also noted in Section 6.3.1 that the headroom consumed by the input transistor, $V_{GS} - V_{TH}$, lowers the conversion gain [Eq. (6.67)]. Along with the above example, these observations point to trade-offs among noise, nonlinearity, gain, and power dissipation in active mixers.

The linearity of active mixers degrades if the switching transistors enter the triode region. To understand this phenomenon, consider the circuit shown in Fig. 6.61, where M_2 is in the triode region while M_3 is still on and in saturation. Note that (1) the load resistors and capacitors establish an output bandwidth commensurate with the IF signal, and (2) the IF signal is uncorrelated with the LO waveform. If both M_2 and M_3 operate in saturation, then the division of I_{RF} between the two transistors is given by their transconductances and is independent of their *drain* voltages.[6] On the other hand, if M_2 is in the triode region, then I_{D2} is a function of the IF voltage at node X, leading to *signal-dependent* current division between M_2 and M_3. To avoid this nonlinearity, M_2 must not enter the triode region so long as M_3 is on and vice versa. Thus, the LO swings cannot be arbitrarily large.

Compression Let us now study gain compression in active mixers. The above effect may manifest itself as the circuit approaches compression. If the output swings become excessively large, the circuit begins to compress at the output rather than at the input, by which we mean the switching devices introduce nonlinearity and hence compression while the input transistor has not reached compression. This phenomenon tends to occur if the gain of the active mixer is relatively high.

Example 6.22

An active mixer exhibits a voltage conversion gain of 10 dB and an input 1-dB compression point of 355 mV_{pp} ($= -5$ dBm). Is it possible that the switching devices contribute compression?

6. We neglect channel-length modulation here.

Example 6.22 *(Continued)*

Solution:

At an input level of -5 dBm, the mixer gain drops to 9 dB, leading to an output differential swing of $355\,\text{mV}_{pp} \times 2.82 \approx 1\ \text{V}_{pp}$. Thus, each output node experiences a *peak* swing of 250 mV; i.e., node X in Fig. 6.61 falls 250 mV below its bias point. If the LO drive is large enough, the switching devices enter the triode region and compress the gain.

The input transistor may introduce compression even if it satisfies the quadratic characteristics of long-channel MOSFETs. This is because, with a large input level, the gate voltage of the device rises while the drain voltage falls, possibly driving it into the triode region. From Fig. 6.51, we can write the RF voltage swing at node P as

$$V_P \approx -g_{m1} R_P V_{RF}, \tag{6.103}$$

where R_P denotes the "average resistance" seen at the common source node of M_2 and M_3.[7] We can approximate R_P as $(1/g_{m2})||(1/g_{m3})$, where g_{m2} and g_{m3} represent the equilibrium transconductances of M_2 and M_3, respectively. In a typical design, $g_{m1}R_P$ is on the order of unity. Thus, in the above example, as the input rises by $355\,\text{mV}/2 = 178\,\text{mV}$ from its bias value, the drain voltage of the input device falls by about 178 mV. If M_1 must not enter the triode region, then the drain-source headroom allocated to M_1 must be 355 mV *higher* than its quiescent overdrive voltage. Note that we did not account for this extra drain voltage swing in Example 6.12. If we had, the conversion gain would have been even lower.

The IP_2 of active mixers is also of great interest. We compute the IP_2 in Section 6.4.3.

Example 6.23

Design a 6-GHz active mixer in 65-nm technology with a bias current of 2 mA from a 1.2-V supply. Assume direct downconversion with a peak single-ended sinusoidal LO swing of 400 mV.

Solution:

The design of the mixer is constrained by the limited voltage headroom. We begin by assigning an overdrive voltage of 300 mV to the input transistor, M_1, and 150 mV to the switching devices, M_2 and M_3 (in equilibrium) (Fig. 6.62). From Eq. (6.64), we obtain a maximum allowable dc drop of about 600 mV for each load resistor, R_D. With a total bias current of 2 mA, we conservatively choose $R_D = 500\,\Omega$. Note that the LO swing well exceeds the voltage necessary to switch M_2 and M_3, forcing I_{D2} or I_{D3} to go from 2 mA to zero in about 5 ps.

(Continues)

7. Since R_P varies periodically, with a frequency equal to $2\omega_{LO}$, we can express its value by a Fourier series and consider the first term as the average resistance.

Example 6.23 *(Continued)*

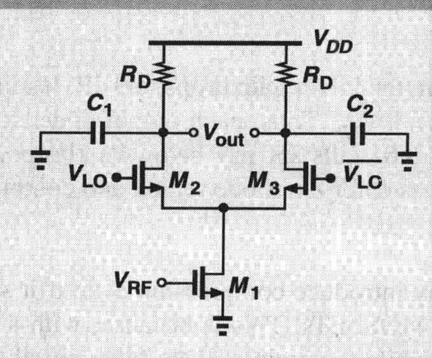


Figure 6.62 *Active mixer design for the 6-GHz band.*

The overdrives chosen above lead to $W_1 = 15\,\mu$m and $W_{2,3} = 20\,\mu$m. According to the g_m-I_D characteristic plotted in Chapter 5 for $W = 10\,\mu$m, g_m reaches approximately 8.5 mS for $I_D = 2\text{ mA} \times (10/15) = 1.33$ mA. Thus, for $W_1 = 15\,\mu$m and $I_{D1} = 2$ mA, we have $g_{m1} = 8.5\text{ mS} \times 1.5 = 12.75\text{ mS} = (78.4\,\Omega)^{-1}$. Capacitors C_1 and C_2 have a value of 2 pF to suppress the LO component at the output (which would otherwise help compress the mixer at the output).

We can now estimate the voltage conversion gain and the noise figure of the mixer. We have

$$A_v = \frac{2}{\pi} g_{m1} R_D \tag{6.104}$$

$$= 4.1\ (= 12.3\text{ dB}). \tag{6.105}$$

To compute the noise figure due to thermal noise, we first estimate the input-referred noise voltage as

$$\overline{V_{n,in}^2} = \pi^2 kT \left(\frac{\gamma}{g_{m1}} + \frac{2}{g_{m1}^2 R_D} \right) \tag{6.106}$$

$$= 4.21 \times 10^{-18}\ \text{V}^2/\text{Hz}, \tag{6.107}$$

where $\gamma \approx 1$. Note that, at a given IF $\neq 0$, this noise results from both the signal band and the image band, ultimately yielding the single-sideband noise figure. We now write the NF with respect to $R_S = 50\,\Omega$ as

$$\text{NF}_{SSB} = 1 + \frac{\overline{V_{n,in}^2}}{4kTR_S} \tag{6.108}$$

$$= 6.1 (= 7.84\text{ dB}). \tag{6.109}$$

The double-sideband NF is 3 dB less.

Example 6.23 (*Continued*)

In the simulation of mixers, we consider nonzero baseband frequencies even for direct-conversion receivers. After all, the RF signal has a finite bandwidth, producing nonzero IF components upon downconversion. For example, a 20-MHz 11a channel occupies a bandwidth of ± 10 MHz in the baseband. Simulations therefore assume an LO frequency, f_{LO}, of, say, 6 GHz, and an input frequency, f_{RF}, of, say, 6.01 GHz. The time-domain simulation must then be long enough to capture a sufficient number of IF cycles for an accurate Fast Fourier Transform (FFT). If the bandwidth at the mixer output nodes permits, we may choose a higher IF to shorten the simulation time.

Figure 6.63 plots the simulated conversion gain of the mixer as a function of the peak input voltage, $V_{in,p}$. Here, $f_{LO} = 6$ GHz, $f_{in} = 5.95$ GHz, and $V_{in,p}$ is increased in each simulation. The uncompressed gain is 10.3 dB, about 2 dB less than our estimate, falling by 1 dB at $V_{in,p} = 170$ mV ($= -5.28$ dBm). Note that LO feedthrough and signal distortion make it difficult to measure the amplitude of the 50-MHz IF in the time domain. For this reason, the FFTs of the input and the output are examined so as to measure the conversion gain.

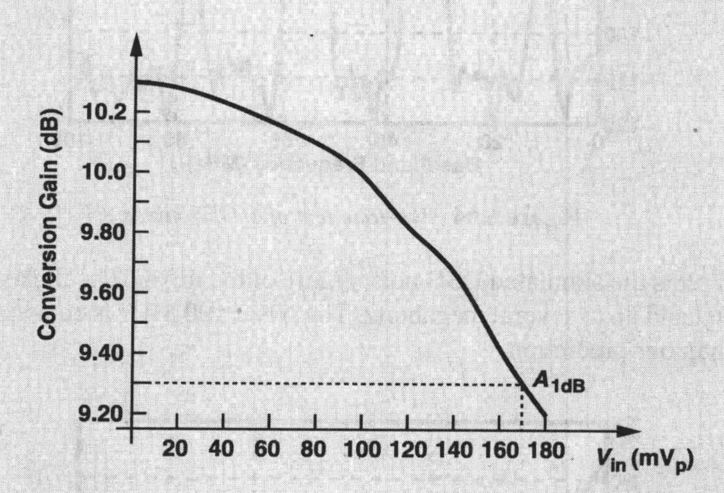


Figure 6.63 *Compression characteristic of 6-GHz mixer.*

Does this mixer design first compress at the input or at the output? As a test, we reduce the load resistors by a factor of 5, scaling the output voltage swings proportionally, and perform the above simulation again. We observe that the gain drops by only 0.5 dB at $V_{in,p} = 170$ mV. Thus, the output port, i.e., the switching transistors, reach compression first.

In order to measure the input IP_3 of the mixer, we apply to the input two sinusoidal voltage sources in series with frequencies equal to 5.945 GHz and 5.967 GHz. The peak amplitude of each tone is chosen after some iteration: if it is too small, the output IM_3 components are corrupted by the FFT noise floor, and if it is too large, the circuit may experience higher-order nonlinearity. We choose a peak amplitude of 40 mV. Figure 6.64 plots the downconverted spectrum, revealing a difference of $\Delta P = 50$ dB between the

(*Continues*)

Example 6.23 *(Continued)*

fundamentals and the IM_3 tones. We divide this value by 2 and by another factor of 20, compute $10^{\Delta P/40} = 17.8$, and multiply the result by the input peak voltage, obtaining $IIP_3 = 711\ mV_p$ ($= +7$ dBm in a 50-Ω system). The IIP_3 is 12.3 dB higher than the input P_{1dB} in this design—perhaps because when the mixer approaches P_{1dB}, its nonlinearity has higher-order terms.

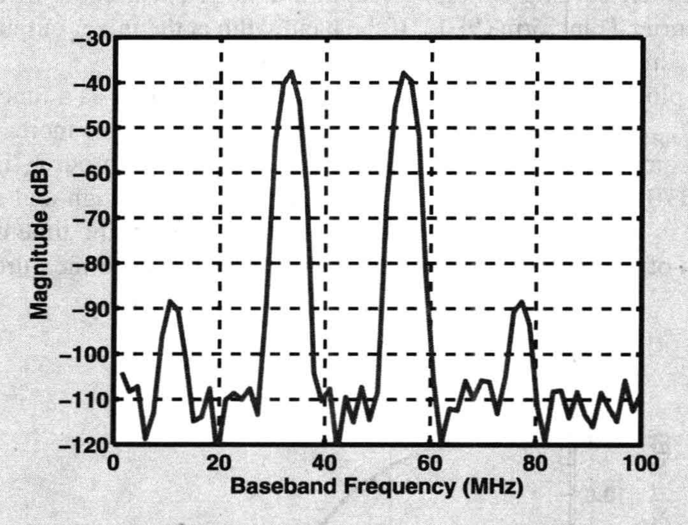


Figure 6.64 *Two-tone test of 6-GHz mixer.*

Figure 6.65 plots the simulated DSB noise figure of the mixer. The flicker noise heavily corrupts the baseband up to several megahertz. The NF at 100 MHz is equal to 5.5 dB, about 0.7 dB higher than our prediction.

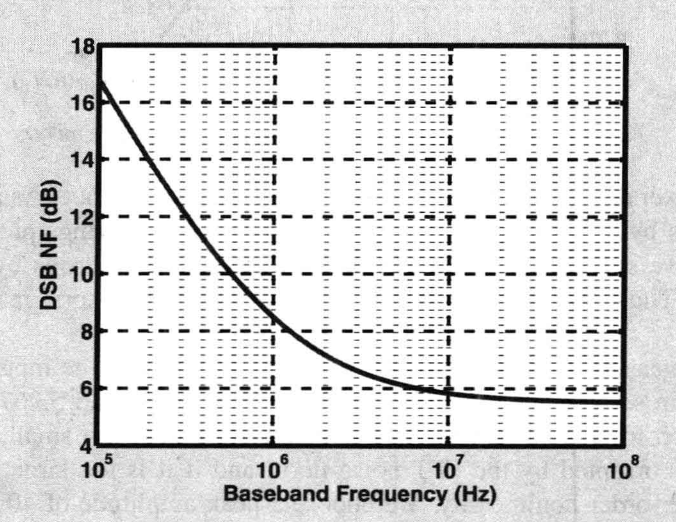


Figure 6.65 *Noise figure of 6-GHz mixer.*

6.4 IMPROVED MIXER TOPOLOGIES

The mixer performance envelope defined by noise, nonlinearity, gain, power dissipation, and voltage headroom must often be pushed beyond the typical scenarios studied thus far in this chapter. For this reason, a multitude of circuit techniques have been introduced to improve the performance of mixers, especially active topologies. In this section, we present some of these techniques.

6.4.1 Active Mixers with Current-Source Helpers

The principal difficulty in the design of active mixers stems from the conflicting requirements between the input transistor current (which must be high enough to meet noise and linearity specifications) and the load resistor current (which must be low enough to allow large resistors and hence a high gain). We therefore surmise that adding current sources ("helpers") in parallel with the load resistors (Fig. 6.66) alleviates this conflict by affording larger resistor values. If $I_{D1} = 2I_0$ and each current source carries a fraction, αI_0, then R_D can be as large as $V_0/[(1-\alpha)I_0]$, where V_0 is the maximum allowable drop across R_D [as formulated by Eq. (6.64)]. Consequently, the voltage conversion gain rises as α increases. For example, if $\alpha = 0.5$, then R_D can be doubled and so can the gain. A higher R_D also reduces its input-referred noise contribution [Eq. (6.85)].

But how about the noise contributed by M_4 and M_5? Assuming that these devices are biased at the edge of saturation, i.e., $|V_{GS} - V_{TH}|_{4,5} = V_0$, we write the noise current of each as $4kT\gamma g_m = 4kT\gamma(2\alpha I_0)/V_0$, multiply it by R_D^2 to obtain the (squared) noise voltage at each output node,[8] and sum the result with the noise of R_D itself:

$$\overline{V_{n,X}^2} = 4kT\gamma\frac{2\alpha I_0}{V_0}R_D^2 + 4kTR_D, \tag{6.110}$$

where the noise due to other parts of the mixer is excluded. Since the voltage conversion gain is proportional to R_D, the above noise power must be normalized to R_D^2 [and eventually

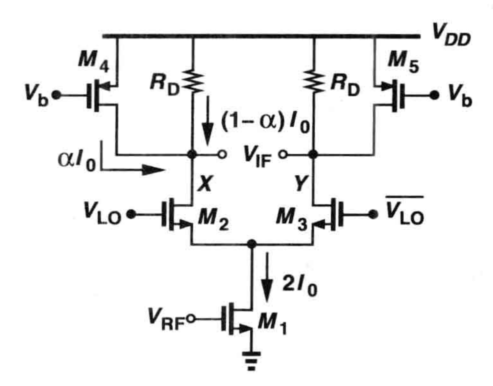


Figure 6.66 *Addition of load current sources to relax headroom constraints.*

8. The output resistance of M_4 and M_5 can be absorbed in R_D for this calculation.

the other factors in Eq. (6.85)]. We thus write

$$\frac{\overline{V_{n,X}^2}}{R_D^2} = 4kT\gamma\frac{2\alpha I_0}{V_0} + \frac{4kT}{R_D} \tag{6.111}$$

$$= 4kT\frac{I_0}{V_0}(2\alpha\gamma + 1 - \alpha) \tag{6.112}$$

$$= 4kT\frac{I_0}{V_0}[(2\gamma - 1)\alpha + 1]. \tag{6.113}$$

Interestingly, the total noise due to each current-source helper and its corresponding load resistor *rises* with α, beginning from $4kTI_0/V_0$ for $\alpha = 0$ and reaching $(4kTI_0/V_0)(2\gamma)$ for $\alpha = 1$.

Example 6.24

Study the flicker noise contribution of M_4 and M_5 in Fig. 6.66.

Solution:

Modeled by a gate-referred voltage, $\overline{V_{n,1/f}^2}$, the flicker noise of each device is multiplied by $g_{m4,5}^2 R_D^2$ as it appears at the output. As with the above derivation, we normalize this result to R_D^2:

$$\frac{\overline{V_{n,X}^2}}{R_D^2} = \overline{V_{n,1/f}^2}\left(\frac{2\alpha I_0}{V_0}\right)^2. \tag{6.114}$$

Since the voltage headroom, V_0, is typically limited to a few hundred millivolts, the helper transistors tend to contribute substantial $1/f$ noise to the output, a serious issue in direct-conversion receivers.

The addition of the helpers in Fig. 6.66 also degrades the linearity. In the calculations leading to Eq. (6.113), we assumed that the helpers operate at the edge of saturation so as to *minimize* their transconductance and hence their noise current, but this bias condition readily drives them into the triode region in the presence of signals. The circuit is therefore likely to compress at the output rather than at the input.

6.4.2 Active Mixers with Enhanced Transconductance

Following the foregoing thought process, we can insert the current-source helper in the *RF path* rather than in the IF path. Depicted in Fig. 6.67 [8], the idea is to provide most of the bias current of M_1 by M_4, thereby reducing the current flowing through the load resistors (and the switching transistors). For example, if $|I_{D4}| = 0.75 I_{D1}$, then R_D and hence the gain can be quadrupled. Moreover, the reduction of the bias current switched by M_2 and M_3 translates to a lower overdrive voltage and more abrupt switching, decreasing ΔT in

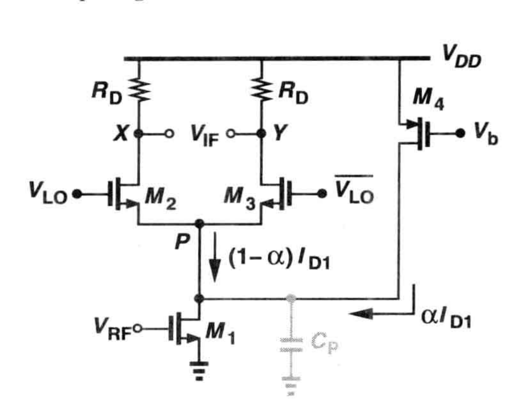


Figure 6.67 *Addition of current source to tail of switching pair.*

Figs. 6.48(a) and 6.58 and lessening the gain and noise effects formulated by Eqs. (6.72) and (6.88). Finally, the output flicker noise falls (Problem 6.10).

The above approach nonetheless faces two issues. First, transistor M_4 contributes additional capacitance to node P, exacerbating the difficulties mentioned earlier. As a smaller bias current is allocated to M_2 and M_3, raising the impedance seen at their source $[\approx 1/(2g_m)]$, C_P "steals" a greater fraction of the RF current generated by M_1, reducing the gain. Second, the noise current of M_4 directly adds to the RF signal. We can readily express the noise currents of M_1 and M_4 as

$$\overline{I^2_{n,M1}} + \overline{I^2_{n,M4}} = 4kT\gamma g_{m1} + 4kT\gamma g_{m4} \tag{6.115}$$

$$= 4kT\gamma \left[\frac{2I_{D1}}{(V_{GS} - V_{TH})_1} + \frac{2\alpha I_{D1}}{|V_{GS} - V_{TH}|_2}\right]. \tag{6.116}$$

Example 6.25

A student eager to minimize the noise of M_4 in the above equation selects $|V_{GS} - V_{TH}|_2 = 0.75$ V with $V_{DD} = 1$ V. Explain the difficulty here.

Solution:

The bias current of M_4 must be carefully defined so as to track that of M_1. Poor matching may "starve" M_2 and M_3, i.e., reduce their bias currents considerably, creating a high impedance at node P and forcing the RF current to ground through C_P. Now, consider the simple current mirror shown in Fig. 6.68. If $|V_{GS} - V_{TH}|_4 = 0.75$ V, then $|V_{GS4}|$ may *exceed* V_{DD}, leaving no headroom for I_{REF}. In other words, $|V_{GS} - V_{TH}|_4$ must be chosen less than $V_{DD} - |V_{GS4}| - V_{IREF}$, where V_{IREF} denotes the minimum acceptable voltage across I_{REF}.

(Continues)

Example 6.25 (*Continued*)

Figure 6.68 *Current mirror voltage limitations.*

In order to suppress the capacitance and noise contribution of M_4 in Fig. 6.68, an inductor can be placed in series with its drain. Illustrated in Fig. 6.69(a) [9], such an arrangement not only enhances the input transconductance but allows the inductor to *resonate* with C_P. Additionally, capacitor C_1 acts as a short at RF, shunting the noise current of M_4 to ground. As a result, most of the RF current produced by M_1 is commutated by M_2 and M_3, and the noise injected by M_2 and M_3 is also reduced (because they switch more abruptly).

In the circuit of Fig. 6.69(a), the inductor parasitics must be managed carefully. First, L_1 contributes some capacitance to node P, equivalently raising C_P. Second, the loss of L_1 translates to a parallel resistance, "wasting" the RF current and adding noise. Depicted in Fig. 6.69(b), this resistance, R_1, must remain much greater than $1/(2g_{m2,3})$ so as to

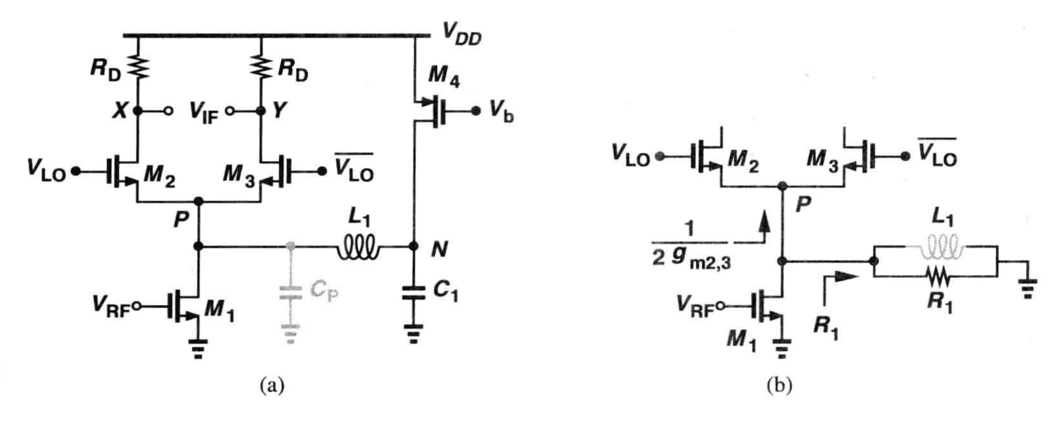


Figure 6.69 *(a) Use of inductive resonance at tail with helper current source, (b) equivalent circuit of inductor.*

negligibly shunt the RF current. Also, its noise current must be much less than that of M_1. Thus, the choice of the inductor is governed by the following conditions:

$$L_1 C_{P,tot} = \frac{1}{\omega_{RF}^2} \tag{6.117}$$

$$R_1 = QL_1\omega_{RF} \gg \frac{1}{g_{m2,3}} \tag{6.118}$$

$$\frac{4kT}{R_1} = \frac{4kT}{QL_1\omega_{RF}} \ll 4kT\gamma g_{m1}, \tag{6.119}$$

where $C_{P,tot}$ includes the capacitance of L_1.

The circuits of Figs. 6.67 and 6.69 suffer from a drawback in deep-submicron technologies: since M_1 is typically a small transistor, it poorly matches the current mirror arrangement that feeds M_4. As a result, the exact current flowing through the switching pair may vary considerably.

Figure 6.70 shows another topology wherein capacitive coupling permits independent bias currents for the input transistor and the switching pair [10]. Here, C_1 acts as a short circuit at RF and L_1 resonates with the parasitics at nodes P and N. Furthermore, the voltage headroom available to M_1 is no longer constrained by $(V_{GS} - V_{TH})_{2,3}$ and the drop across the load resistors. In a typical design, I_{D1}/I_0 may fall in the range of 3 to 5 for optimum performance. Note that if I_0 is excessively low, the switching pair does not absorb all of the RF current. Another important attribute is that, as formulated by Eq. (6.97), a smaller I_0 leads to lower flicker noise at the output.

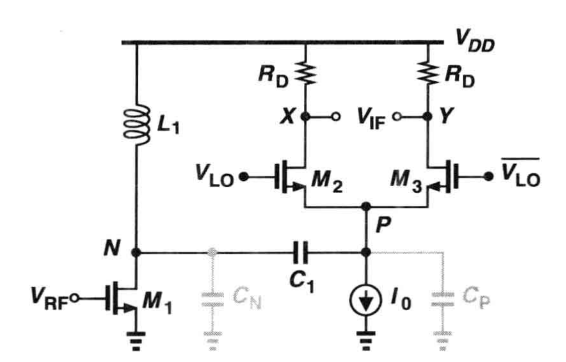


Figure 6.70 *Active mixer using capacitive coupling with resonance.*

6.4.3 Active Mixers with High IP_2

As explained in Chapter 4, the second intercept point becomes critical in direct-conversion and low-IF receivers as it signifies the corruption introduced by the beating of two interferers or envelope demodulation of one interferer. We also noted that capacitive

coupling between the LNA and the mixer removes the low-frequency beat, making the mixer the bottleneck. Thus, a great deal of effort has been expended on high-IP_2 mixers.

It is instructive to compute the IP_2 of a single-balanced mixer in the presence of asymmetries. (Recall from Chapter 4 that a symmetric mixer has an infinite IP_2.) Let us begin with the circuit of Fig. 6.71(a), where V_{OS} denotes the offset voltage associated with M_2 and M_3. We wish to compute the fraction of I_{SS} that flows to the output *without* frequency translation. As with the flicker noise calculations in Section 6.3.2, we assume LO and $\overline{\text{LO}}$ exhibit a finite slope but M_2 and M_3 switch instantaneously, i.e., they switch the tail current according to the *sign* of $V_A - V_B$.

As shown in Fig. 6.71(b), the vertical shift of V_{LO} displaces the consecutive crossings of LO and $\overline{\text{LO}}$ by $\pm\Delta T$, where $\Delta T = V_{OS}/S_{LO}$ and S_{LO} denotes the differential slope of the LO $(= 2V_{p,LO}\omega_{LO})$. This forces M_2 to remain on for $T_{LO}/2 + 2\Delta T$ seconds and M_3 for $T_{LO}/2 - 2\Delta T$ seconds. It follows from Fig. 6.71(c) that the differential output current, $I_{D2} - I_{D3}$ contains a dc component equal to $(4\Delta T/T_{LO})I_{SS} = V_{OS}I_{SS}/(\pi V_{p,LO})$, and the differential output voltage a dc component equal to $V_{OS}I_{SS}R_D/(\pi V_{p,LO})$. As expected,

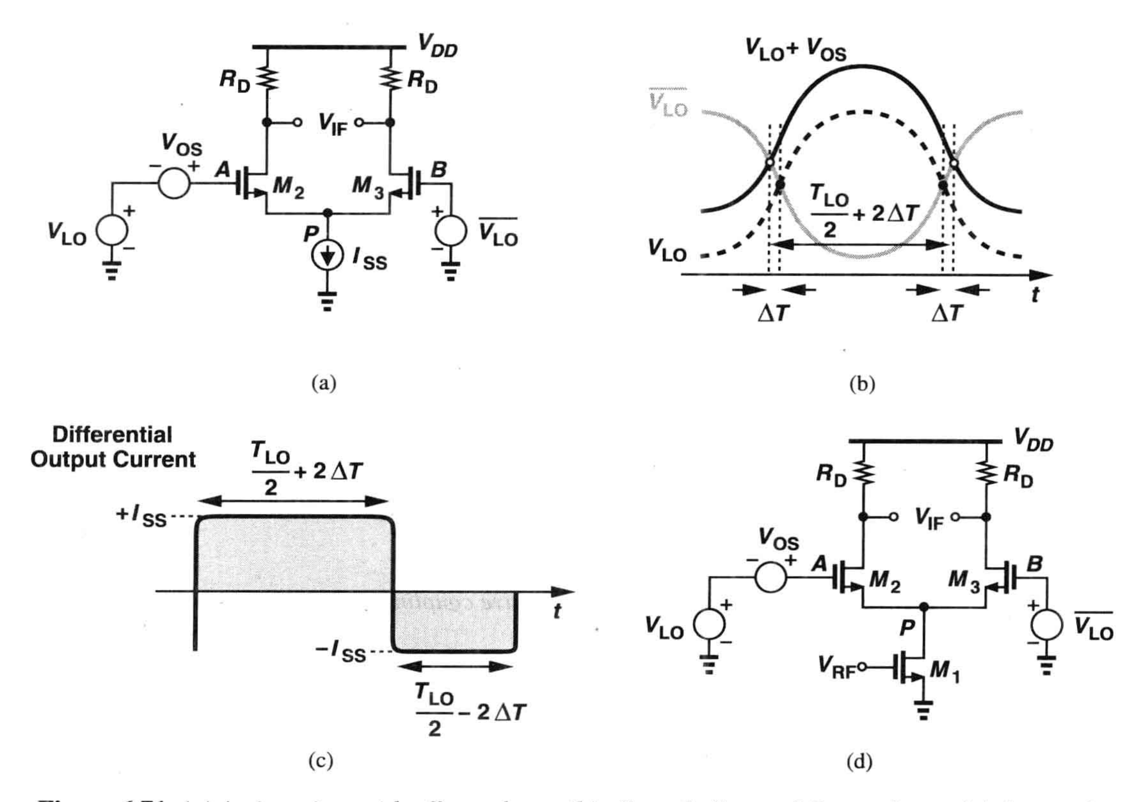


Figure 6.71 *(a) Active mixer with offset voltage, (b) effect of offset on LO waveforms, (c) duty cycle distortion of drain currents, (d) circuit for IP_2 computation .*

this result agrees with Eq. (6.97) because the offset can be considered a very slow noise component.

An interesting observation offered by the output $1/f$ noise and offset equations is as follows. If the bias current of the switching pair is reduced but that of the input transconductor is not, then the performance improves because the gain does not change but the output $1/f$ noise and offset fall. For example, the current helpers described in the previous section prove useful here.

We now replace I_{SS} with a transconductor device as depicted in Fig. 6.71(d) and assume

$$V_{RF} = V_m \cos \omega_1 t + V_m \cos \omega_2 t + V_{GS0}, \tag{6.120}$$

where V_{GS0} is the bias gate-source voltage of M_1. With a square-law device, the IM_2 product emerges in the current of M_1 as

$$I_{IM2} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}V_m^2 \cos(\omega_1 - \omega_2)t. \tag{6.121}$$

Multiplying this quantity by $V_{OS}R_D/(\pi V_{p,LO})$ yields the direct feedthrough to the output:

$$V_{IM2,out} = \left[\frac{1}{2}\mu_n C_{ox}\frac{W}{L}V_m^2 \cos(\omega_1 - \omega_2)t\right]\frac{V_{OS}R_D}{\pi V_{p,LO}}. \tag{6.122}$$

To calculate the IP_2, the value of V_m must be raised until the amplitude of $V_{IM2,out}$ becomes equal to the amplitude of the main *downconverted* components. This amplitude is simply given by $(2/\pi)g_{m1}R_D V_m$. Thus,

$$\frac{1}{2}\mu_n C_{ox}\frac{W}{L}V_{IIP2}^2\frac{V_{OS}R_D}{\pi V_{p,LO}} = \frac{2}{\pi}g_{m1}R_D V_{IIP2}. \tag{6.123}$$

Writing g_{m1} as $\mu_n C_{ox}(W/L)(V_{GS} - V_{TH})_1$, we finally obtain

$$V_{IIP2} = 4(V_{GS} - V_{TH})_1\frac{V_{p,LO}}{V_{OS}}. \tag{6.124}$$

For example, if $(V_{GS} - V_{TH})_1 = 250\,\text{mV}$, $V_{p,LO} = 300\,\text{mV}$, and $V_{OS} = 10\,\text{mV}$, then $V_{IIP2} = 30\ V_P$ ($= 39.5\,\text{dBm}$ in a 50-Ω system). Other IP_2 mechanisms are described in [12].

The foregoing analysis also applies to asymmetries in the LO waveforms that would arise from mismatches within the LO circuitry and its buffer. If the duty cycle is denoted by $(T_{LO}/2 - \Delta T)/T_{LO}$ (e.g., 48%), then the dc component in $I_{D1} - I_{D2}$ is equal to $(2\Delta T/T_{LO})I_{SS}$, yielding an average of $(2\Delta T/T_{LO})I_{SS}R_D$ at the output. We therefore replace I_{SS} with the IM_2 component given by Eq. (6.121), arriving at

$$V_{IM2,out} = \left[\frac{1}{2}\mu_n C_{ox}\frac{W}{L}V_m^2 \cos(\omega_1 - \omega_2)t\right]\frac{2\Delta T}{T_{LO}}R_D. \tag{6.125}$$

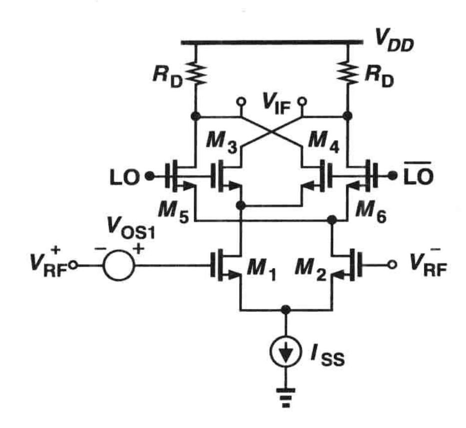


Figure 6.72 *Input offset in a double-balanced mixer.*

Equating the amplitude of this component to $(2/\pi)g_{m1}R_DV_m$ and substituting $\mu_nC_{ox}(W/L)(V_{GS} - V_{TH})_1$ for g_{m1}, we have

$$V_{IIP2} = \frac{2T_{LO}}{\pi \Delta T}(V_{GS} - V_{TH})_1. \tag{6.126}$$

For example, a duty cycle of 48% along with $(V_{GS} - V_{TH})_1 = 250\,\text{mV}$ gives rise to $V_{IIP2} = 7.96\ V_P$ (= 28 dBm in a 50-Ω system).

In order to raise the IP_2, the input transconductor of an active mixer can be realized in differential form, leading to a double-balanced topology. Shown in Fig. 6.72, such a circuit produces a finite IM_2 product only as a result of *mismatches* between M_1 and M_2. We quantify this effect in the following example. Note that, unlike the previous double-balanced mixers, this circuit employs a tail current source.

Example 6.26

Assuming square-law devices, determine the IM_2 product generated by M_1 and M_2 in Fig. 6.72 if the two transistors suffer from an offset voltage of V_{OS1}.

Solution:

For an RF differential voltage, ΔV_{in}, the differential output current can be expressed as

$$I_{D1} - I_{D2} = \frac{1}{2}\mu_nC_{ox}\frac{W}{L}(\Delta V_{in} - V_{OS1})\sqrt{\frac{4I_{SS}}{\mu_nC_{ox}(W/L)} - (\Delta V_{in} - V_{OS1})^2}. \tag{6.127}$$

Assuming that the second term under the square root is much less than the first, we write $\sqrt{1-\varepsilon} \approx 1 - \varepsilon/2$:

$$I_{D1} - I_{D2} \approx \sqrt{\mu_nC_{ox}\frac{W}{L}I_{SS}}\left[\Delta V_{in} - V_{OS1} - \frac{\mu_nC_{ox}(W/L)}{8I_{SS}}(\Delta V_{in} - V_{OS1})^3\right]. \tag{6.128}$$

Example 6.26 *(Continued)*

The cubic term in the square brackets produces an IM_2 component if $\Delta V_{in} = V_m \cos\omega_1 t + V_m \cos\omega_2 t$ because the term $3\Delta V_{in}^2 V_{OS1}$ leads to the cross product of the two sinusoids:

$$V_{IM2} = \frac{3[\mu_n C_{ox}(W/L)]^{3/2}}{8\sqrt{I_{SS}}} V_m^2 V_{OS1} \cos(\omega_1 - \omega_2)t \tag{6.129}$$

$$= \frac{3I_{SS}}{8(V_{GS} - V_{TH})_{eq}^3} V_m^2 V_{OS1} \cos(\omega_1 - \omega_2)t, \tag{6.130}$$

where $(V_{GS} - V_{TH})_{eq}$ represents the equilibrium overdrive of each transistor. Of course, only a small fraction of this component appears at the output of the mixer. For example, if only the offset of the switching quad, V_{OS2}, is considered,[9] then the IM_2 amplitude must be multiplied by $V_{OS2}R_D/(\pi V_{p,LO})$, yielding an IIP_2 of

$$V_{IIP2} = \frac{16(V_{GS} - V_{TH})_{eq}^2 V_{p,LO}}{3V_{OS1}V_{OS2}}. \tag{6.131}$$

For example, if $(V_{GS} - V_{TH})_{eq} = 250$ mV, $V_{p,LO} = 300$ mV, and $V_{OS1} = V_{OS2} = 10$ mV, then $V_{IIP2} = 1000\ V_p$ (= +70 dBm in a 50-Ω system).

While improving the IP_2 significantly, the use of a differential pair in Fig. 6.72 *degrades* the IP_3. As formulated in Chapter 5, a quasi-differential pair (with the sources held at ac ground) exhibits a higher IP_3. We now repeat the calculations leading to Eq. (6.131) for such a mixer (Fig. 6.73), noting that the input pair now has poor common-mode

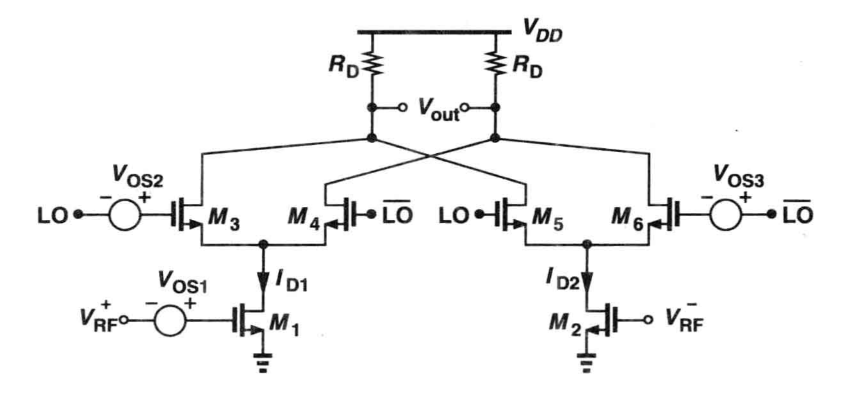


Figure 6.73 *Effect of offsets in a double-balanced mixer using a quasi-differential input pair.*

9. In this case, V_{OS2} represents the *difference* between the offsets of M_3–M_4 and M_5–M_6.

rejection. Let us apply $V_{RF}^+ = V_m \cos \omega_1 t + V_m \cos \omega_2 t + V_{GS0}$ and $V_{RF}^- = -V_m \cos \omega_1 t - V_m \cos \omega_2 t + V_{GS0}$, obtaining

$$I_{D1} = \frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right)_1 (V_m \cos \omega_1 t + V_m \cos \omega_2 t + V_{OS1} + V_{GS0} - V_{TH})^2 \quad (6.132)$$

$$I_{D2} = \frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right)_2 (V_m \cos \omega_1 t + V_m \cos \omega_2 t + V_{GS0} - V_{TH})^2. \quad (6.133)$$

While *independent* of V_{OS1}, the low-frequency beat in I_{D1} is multiplied by a factor of $V_{OS2}R_D/(\pi V_{p,LO})$ and that in I_{D2} by $V_{OS3}R_D/(\pi V_{p,LO})$. Here, V_{OS2} and V_{OS3} denote the offsets of M_3–M_4 and M_5–M_6, respectively. The output thus exhibits an IM_2 component given by

$$V_{IM2,out} = \left[\frac{1}{2}\mu_n C_{ox}\frac{W}{L}V_m^2 \cos(\omega_1 - \omega_2)t\right]\frac{R_D}{\pi V_{p,LO}}(V_{OS2} + V_{OS3}). \quad (6.134)$$

Noting that the output amplitude of each fundamental is equal to $(2/\pi)2V_m g_{m1} R_D$ and that $g_{m1} = \mu_n C_{ox}(W/L)_1(V_{GS0} - V_{TH})$, we have

$$V_{IIP2} = \frac{8(V_{GS0} - V_{TH})}{V_{OS2} + V_{OS3}}V_{p,LO}. \quad (6.135)$$

For example, if $V_{GS} - V_{TH} = 250\,\text{mV}$, $V_{p,LO} = 300\,\text{mV}$, and $V_{OS2} = V_{OS3} = +10\,\text{mV}$, then $V_{IIP2} = 30\,\text{V}_p$ ($= +39.5$ dBm in a 50-Ω system). Comparison of the IIP_2's obtained for the differential and quasi-differential mixers indicates that the latter is much inferior, revealing a trade-off between IP_2 and IP_3.

We have thus far considered one mechanism leading to a finite IP_2: the passage of the *low-frequency* beat through the mixer's switching devices. On the other hand, even with no even-order distortion in the transconductor, it is still possible to observe a finite low-frequency beat at the output if (a) the switching devices (or the LO waveforms) exhibit asymmetry *and* (b) a finite capacitance appears at the common source node of the switching devices [11, 12]. In this case, two interferers, $V_m \cos \omega_1 t + V_m \cos \omega_2 t$, arriving at the common source node experience nonlinearity and mixing with the LO harmonics, thereby generating a component at $\omega_1 - \omega_2$ after downconversion. The details of this mechanism are described in [11, 12].

While conceived for noise and gain optimization reasons, the mixer topology in Fig. 6.70 also exhibits a high IP_2. The high-pass filter consisting of L_1, C_1, and the resistance seen at node P suppresses low-frequency beats generated by the even-order distortion in M_1. From the equivalent circuit shown in Fig. 6.74, we have

$$\frac{I_m}{I_{beat}} = \frac{L_1 s}{L_1 s + \dfrac{1}{C_1 s} + \dfrac{1}{2g_m}} \quad (6.136)$$

$$= \frac{L_1 C_1 s^2}{L_1 C_1 s^2 + \dfrac{C_1 s}{2g_m} + 1}. \quad (6.137)$$

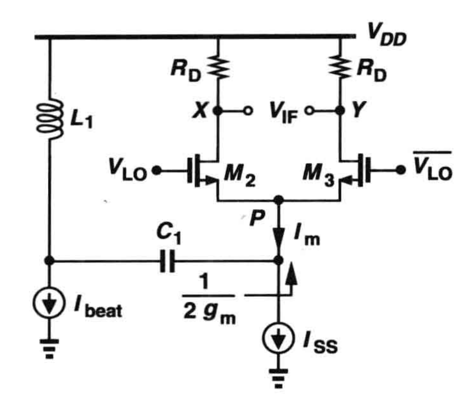


Figure 6.74 *Effect of low-frequency beat in a mixer using capacitive coupling and resonance.*

At low frequencies, this result can be approximated as

$$\frac{I_m}{I_{beat}} \approx L_1 C_1 s^2, \tag{6.138}$$

revealing a high attenuation.

Another approach to raising the IP_2 is to degenerate the transconductor *capacitively*. As illustrated in Fig. 6.75 [10], the degeneration capacitor, C_d, acts as a short circuit at RF but nearly an open circuit at the low-frequency beat components. Expressing the transconductance of the input stage as

$$G_m = \frac{g_{m1}}{1 + \frac{g_{m1}}{C_d s}} \tag{6.139}$$

$$= \frac{g_{m1} C_d s}{C_d s + g_{m1}}, \tag{6.140}$$

we recognize that the gain at low frequencies falls in proportion to $C_d s$, making M_1 *incapable* of generating second-order intermodulation components.

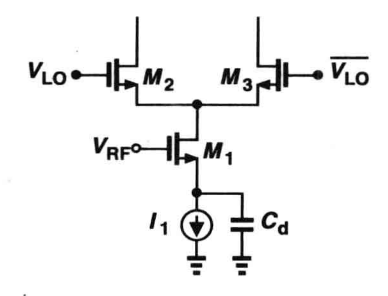


Figure 6.75 *Effect of capacitive degeneration on IP_2.*

Example 6.27

The mixer of Fig. 6.75 is designed for a 900-MHz GSM system. What is the worst-case attenuation that capacitive degeneration provides for IM_2 products that would otherwise be generated by M_1? Assume a low-IF receiver (Chapter 4).

Solution:

We must first determine the worst-case scenario. We may surmise that the *highest* beat frequency experiences the *least* attenuation, thereby creating the largest IM_2 product. As depicted in Fig. 6.76(a), this situation arises if the two interferers remain *within* the GSM band (so that they are not attenuated by the front-end filter) but as far from each other as possible, i.e., at a frequency difference of 25 MHz. Let us assume that the pole frequency, g_m/C_d, is around 900 MHz. The IM_2 product therefore falls at 25 MHz and, therefore, experiences an attenuation of roughly 900 MHz/25 MHz = 36 ($\approx$31 dB) by capacitive degeneration. However, in a low-IF receiver, the downconverted 200-kHz GSM channel is located near zero frequency. Thus, this case proves irrelevant.

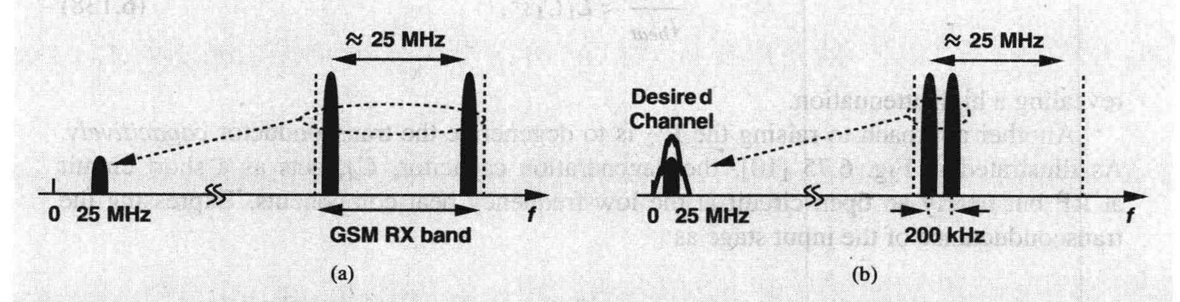


Figure 6.76 *Beat generation from (a) two blockers near the edges of GSM band, (b) two closely-spaced blockers in GSM band.*

From the above study, we seek two interferers that bear a frequency difference of 200 kHz (i.e., adjacent channels). As shown in Fig. 6.76(b), we place the adjacent interferers near the edge of the GSM band. Located at a center frequency of 200 kHz, the beat experiences an attenuation of roughly 935 MHz/200 kHz = 4, 675 $\approx$ 73 dB. It follows that very high IP_2's can be obtained for low-IF 900-MHz GSM receivers.

As mentioned earlier, even with capacitive coupling between the transconductor stage and the switching devices, the capacitance at the common source node of the switching pair ultimately limits the IP_2 (if the offset of the switching pair is considered). We therefore expect a higher IP_2 if an inductor resonates with this capacitance. Figure 6.77 shows a double-balanced mixer employing both capacitive degeneration and resonance to achieve an IP_2 of + 78 dBm [11].

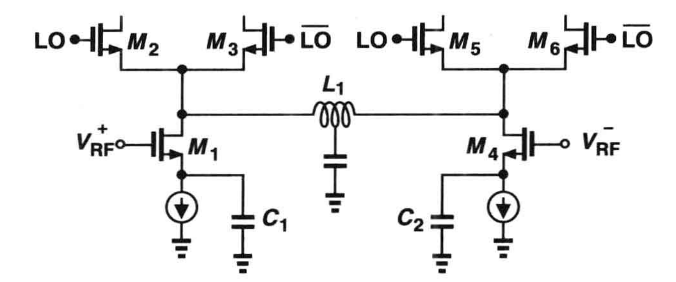


Figure 6.77 *Use of inductor at sources of switching quad to raise IP_2.*

6.4.4 Active Mixers with Low Flicker Noise

Our study of noise in Section 6.3.2 revealed that the downconverted flicker noise of the switching devices is proportional to their bias current and the parasitic capacitance at their common source node. Since these trends also hold for the IP_2 of active mixers, we postulate that the techniques described in Section 6.4.3 for raising the IP_2 lower flicker noise as well. In particular, the circuit topologies in Figs. 6.69 and 6.74 both allow a lower bias current for the switching pair *and* cancel the tail capacitance by the inductor. This approach, however, demands two inductors (one for each quadrature mixer), complicating the layout and routing.

Let us return to the helper idea shown in Fig. 6.67 and ask, is it possible to turn on the helper only at the time when it is needed? In other words, can we turn on the PMOS current source only at the zero crossings of the LO so that it lowers the bias current of the switching devices and hence the effect of their flicker noise [13]? In such a scheme, the helper itself would inject only *common-mode* noise because it turns on only when the switching pair is in equilibrium.

Figure 6.78 depicts our first attempt in realizing this concept. Since large LO swings produce a reasonable voltage swing at node P at $2\omega_{LO}$, the diode-connected transistor turns on when LO and $\overline{LO}$ cross and V_P falls. As LO or $\overline{LO}$ rises, so does V_P, turning M_H off. Thus, M_H can provide most of the bias current of M_1 near the crossing points of LO and $\overline{LO}$ while injecting minimal noise for the rest of the period.

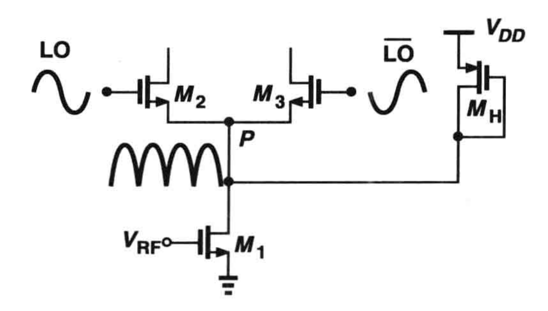


Figure 6.78 *Use of a diode-connected device to reduce switching pair current.*

Unfortunately, the diode-connected transistor in Fig. 6.78 does not turn off abruptly as LO and $\overline{LO}$ depart from their crossing point. Consequently, M_H continues to present a low impedance at node P, shunting the RF current to ac ground. This issue can be resolved in a double-balanced mixer by reconfiguring the diode-connected devices as a *cross-coupled* pair [13]. As illustrated in Fig. 6.79 [13], M_{H1} and M_{H2} turn on and off simultaneously because V_P and V_Q vary identically—as if M_{H1} and M_{H2} were diode-connected devices. Thus, these two transistors provide most of the bias currents of M_1 and M_4 at the crossing points of LO and $\overline{\text{LO}}$. On the other hand, as far as the *differential* RF current of M_1 and M_4 is concerned, the cross-coupled pair acts as a *negative* resistance (Chapter 8), partially cancelling the positive resistance presented by the switching pairs at P and Q. Thus, M_{H1} and M_{H2} do not shunt the RF current.

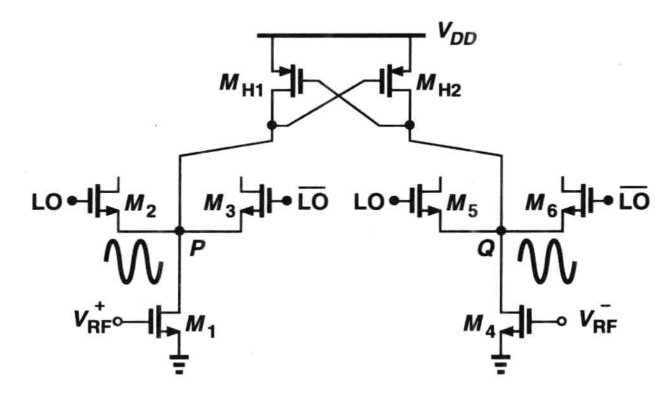


Figure 6.79 *Use of cross-coupled pair to reduce current of switching quad.*

The circuit of Fig. 6.79 nonetheless requires large LO swings to ensure that V_P and V_Q rise rapidly and sufficiently so as to turn off M_{H1} and M_{H2}.[10] Otherwise, these two devices continue to inject differential noise for part of the period. Another drawback of this technique is that it does not lend itself to single-balanced mixers.

Example 6.28

The positive feedback around M_{H1} and M_{H2} in Fig. 6.79 may cause latchup, i.e., a slight imbalance between the two sides may pull P (or Q) toward V_{DD}, turning M_{H2} (or M_{H1}) off. Derive the condition necessary to avoid latchup.

Solution:

The impedance presented by the switching pairs at P and Q is at its *highest* value when either transistor in each differential pair is off (why?). Shown in Fig. 6.80 is the resulting worst case. For a symmetric circuit, the loop gain is equal to $(g_{mH}/g_{m2,5})^2$, where g_{mH}

10. Note that M_{H1} and M_{H2} do not help the switching of the differential pairs because the $2\omega_{LO}$ waveforms at P and Q are *identical* (rather than differential).

Example 6.28 *(Continued)*

represents the transconductance of M_{H1} and M_{H2}. To avoid latchup, we must ensure that

$$\left(\frac{g_{mH}}{g_{m2}}\right)^2 < 1. \tag{6.141}$$

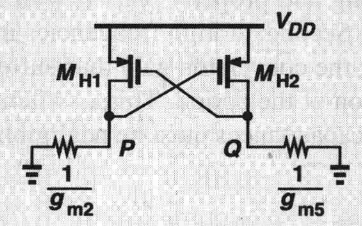


Figure 6.80 *Equivalent circuit for latchup calculation.*

The notion of reducing the current through the switching devices at the crossing points of LO and $\overline{\text{LO}}$ can alternatively be realized by turning off the *transconductor* momentarily [14]. Consider the circuit shown in Fig. 6.81(a), where switch S_1 is driven by a waveform having a frequency of $2f_{LO}$ but a duty cycle of, say, 80%. As depicted in Fig. 6.81(b), S_1

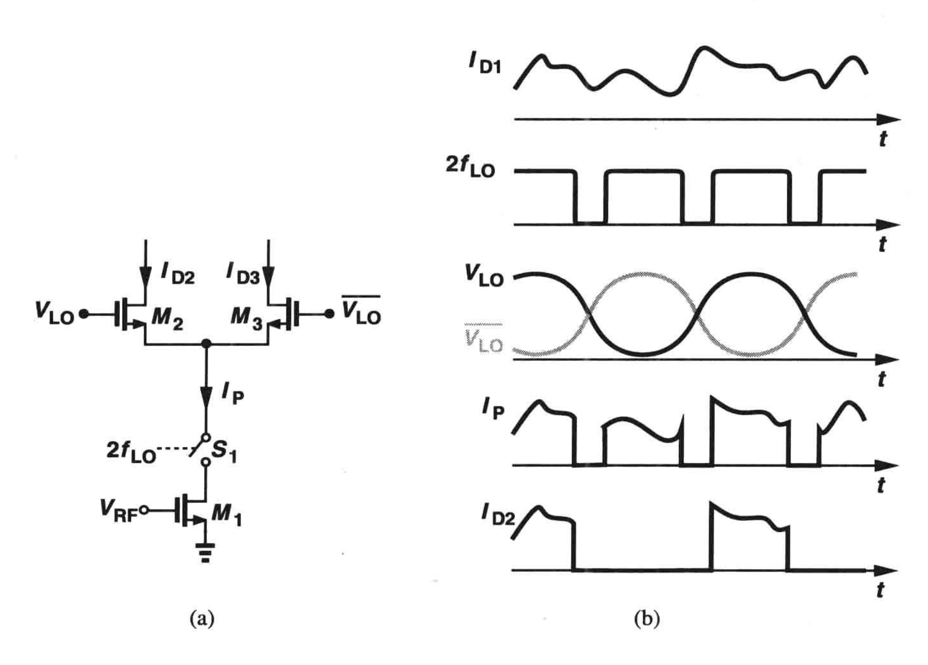


Figure 6.81 *(a) Use of a switch to turn off the switching pair near LO zero crossings, (b) circuit waveforms.*

briefly turns the transconductor off *twice* per LO period. Thus, if the crossing points of LO and $\overline{LO}$ are chosen to fall at the times when I_P is zero, then the flicker noise of M_2 and M_3 is heavily attenuated. Moreover, M_2 and M_3 inject no thermal noise to the output near the equilibrium. The concept can be extended to quadrature double-balanced mixers [14]. In Problem 6.12, we decide whether this circuit can also be viewed as a differential pair whose current is modulated (chopped) at a rate of $2f_{LO}$.

The above approach entails a number of issues. First, the turn-off time of the transconductor must be sufficiently long and properly *phased* with respect to LO and $\overline{LO}$ so that it *encloses* the LO transitions. Second, at high frequencies it becomes difficult to generate $2f_{LO}$ with such narrow pulses; the conversion gain thus suffers because the transconductor remains off for a greater portion of the period. Third, switch S_1 in Fig. 6.81 does consume some voltage headroom if its capacitances must be negligible.

6.5 UPCONVERSION MIXERS

The transmitter architectures studied in Chapter 4 employ upconversion mixers to translate the baseband spectrum to the carrier frequency in one or two steps. In this section, we deal with the design of such mixers.

6.5.1 Performance Requirements

Consider the generic transmitter shown in Fig. 6.82. The design of the TX circuitry typically begins with the PA and moves backward; the PA is designed to deliver the specified power to the antenna while satisfying certain linearity requirements (in terms of the adjacent-channel power or 1-dB compression point). The PA therefore presents a certain input capacitance and, owing to its moderate gain, demands a certain input swing. Thus, the upconversion mixers must (1) translate the baseband spectrum to a *high* output frequency (unlike downconversion mixers) while providing sufficient gain, (2) drive the input capacitance of the PA, (3) deliver the necessary swing to the PA input, and (4) *not* limit the linearity of the TX. In addition, as studied in Chapter 4, dc offsets in upconversion mixers translate to carrier feedthrough and must be minimized.

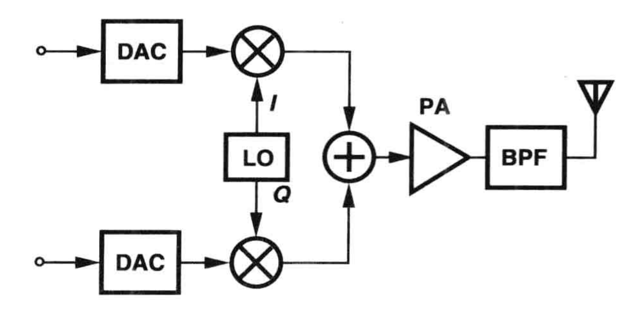


Figure 6.82 *Generic transmitter.*

Example 6.29

Explain the pros and cons of placing a buffer before the PA in Fig. 6.82.

Solution:

The buffer relaxes the drive and perhaps output swing requirements of the upconverter. However, it may contribute significant nonlinearity. For this reason, it is desirable to minimize the number of stages between the mixers and the antenna.

The interface between the mixers and the PA entails another critical issue. Since the baseband and mixer circuits are typically realized in differential form, and since the antenna is typically single-ended, the designer must decide at what point and how the differential output of the mixers must be converted to a single-ended signal. As explained in Chapter 5, this operation presents many difficulties.

The noise requirement of upconversion mixers is generally much more relaxed than that of downconversion mixers. As studied in Problem 6.13, this is true even in GSM, wherein the amplified noise of the upconversion mixers in the receive band must meet certain specifications (Chapter 4).

The interface between the baseband DACs and the upconversion mixers in Fig. 6.82 also imposes another constraint on the design. Recall from Chapter 4 that high-pass filtering of the baseband signal introduces intersymbol interference. Thus, the DACs must be directly coupled to the mixers to avoid a notch in the signal spectrum.[11] As seen below, this issue dictates that the bias conditions in the upconversion mixers be relatively independent of the output common-mode level of the DACs.

6.5.2 Upconversion Mixer Topologies

Passive Mixers The superior linearity of passive mixers makes them attractive for upconversion as well. We wish to construct a quadrature upconverter using passive mixers.

Our study of downconversion mixers has revealed that single-balanced sampling topologies provide a conversion gain that is about 5.5 dB higher than their return-to-zero counterparts. Is this true for upconversion, too? Consider a low-frequency baseband sinusoid applied to a sampling mixer (Fig. 6.83). The output appears to contain mostly the input waveform and *little* high-frequency energy. To quantify our intuition, we return to the constituent waveforms, $y_1(t)$ and $y_2(t)$, given by Eqs. (6.12) and (6.16), respectively, and reexamine them for upconversion, assuming that $x(t)$ is a baseband signal. The component of interest in $Y_1(f)$ still occurs at $k = \pm 1$ and is given by

$$Y_1(f)|_{k=\pm 1} = \frac{X(f - f_{LO})}{j\pi} - \frac{X(f + f_{LO})}{j\pi}. \tag{6.142}$$

11. In reality, each DAC is followed by a low-pass filter to suppress the DAC's high-frequency output components.

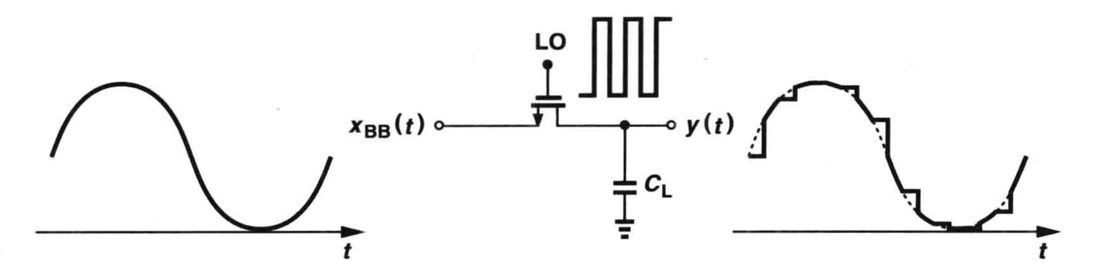


Figure 6.83 *Sampling mixer for upconversion.*

For $Y_2(f)$, we must also set k to ± 1:

$$Y_2(f)|_{k=\pm 1} = \frac{1}{T_{LO}}[-X(f-f_{LO}) + X(f+f_{LO})]\left[\frac{1}{j\omega}(1 - e^{-j\omega T_{LO}/2})\right]. \tag{6.143}$$

However, the term in the second set of brackets must be evaluated at the *upconverted* frequency. If $\omega = \omega_{LO} + \omega_{BB}$, where ω_{BB} denotes the baseband frequency, then $\exp(-j\omega T_{LO}/2) = \exp(-j\pi)\exp(-j\omega_{BB}T_{LO}/2)$, which, for $\omega_{BB} \ll 2f_{LO}$, reduces to $-(1 - j\omega_{BB}T_{LO}/2)$. Similarly, if $\omega = -\omega_{LO} - \omega_{BB}$, then $\exp(-j\omega T_{LO}/2) \approx -(1 + j\omega_{BB}T_{LO}/2)$. Adding $Y_1(f)$ and $Y_2(f)$ gives

$$[Y_1(f) + Y_2(f)]_{k=\pm 1} \approx \frac{\omega_{BB}}{\omega_{LO} + \omega_{BB}}\left[\left(\frac{1}{j\pi} + \frac{1}{2}\right)X(f - f_{LO}) + \left(-\frac{1}{j\pi} + \frac{1}{2}\right)X(f + f_{LO})\right], \tag{6.144}$$

indicating that the upconverted output amplitude is proportional to $\omega_{BB}/(\omega_{LO} + \omega_{BB}) \approx \omega_{BB}/\omega_{LO}$. Thus, such a mixer is not suited to upconversion.

In Problem 6.14, we study a *return-to-zero* mixer for upconversion and show that its conversion gain is still equal to $2/\pi$ (for a single-balanced topology). Similarly, from Example 6.8, a double-balanced passive mixer exhibits a gain of $2/\pi$. Depicted in Fig. 6.84(a), such a topology is more relevant to TX design than single-balanced structures because the baseband waveforms are typically available in differential form. We thus focus on double-balanced mixers here.

While simple and quite linear, the circuit of Fig. 6.84(a) must deal with a number of issues. First, the bandwidth at nodes X and Y must accommodate the upconverted signal frequency so as to avoid additional loss. This bandwidth is determined by the on-resistance of the switches (R_{on}), their capacitance contributions to the output nodes, and the input capacitance of the next stage (C_{in}). Wider switches increase the bandwidth up to the point where their capacitances overwhelm C_{in}, but they also present a greater capacitance at the LO ports.

It is possible to null the capacitance at nodes X and Y by means of resonance. As illustrated in Fig. 6.84(b) [15], inductor L_1 resonates with the total capacitance at X and Y, and its value is chosen to yield

$$\omega_{IF} = \frac{1}{\sqrt{\frac{L_1}{2}(C_{X,Y} + C_{in})}}, \tag{6.145}$$

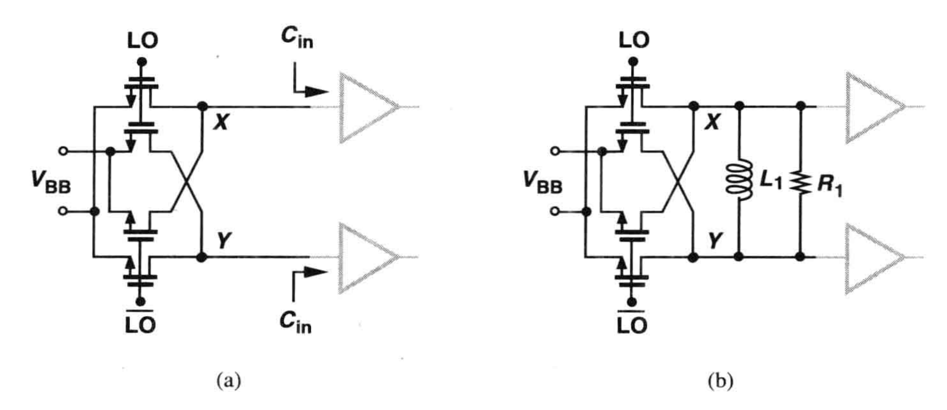


Figure 6.84 *(a) Double-balanced upconversion passive mixer, (b) use of resonance to increase bandwidth.*

where $C_{X,Y}$ denotes the capacitances contributed by the switches at X or Y. At resonance, the mixers are loaded by the parallel equivalent resistance of the inductor, $R_1 = QL_1\omega_{IF}$. Thus, we require that $2R_{on} \ll R_1$ to avoid additional loss. This technique becomes necessary only at very high frequencies, e.g., at 50 GHz and above.

The second issue relates to the use of passive mixers in a quadrature upconverter, where the outputs of two mixers must be summed. Unfortunately, passive mixers sense and produce *voltages*, making direct summation difficult. We therefore convert each output to current, sum the currents, and convert the result to voltage. Figure 6.85(a) depicts such an arrangement. Here, the quasi-differential pairs M_1–M_2 and M_3–M_4 perform V/I conversion, and the load resistors, I/V conversion. This circuit can provide gain while lending itself to low supply voltages. The grounded sources of M_1–M_4 also yield a relatively high linearity.[12]

A drawback of the above topology is that its bias point is sensitive to the input common-mode level, i.e., the output CM level of the preceding DAC. As shown in Fig. 6.85(b), I_{D1} depends on V_{BB} and varies significantly with process and temperature. For this reason, we

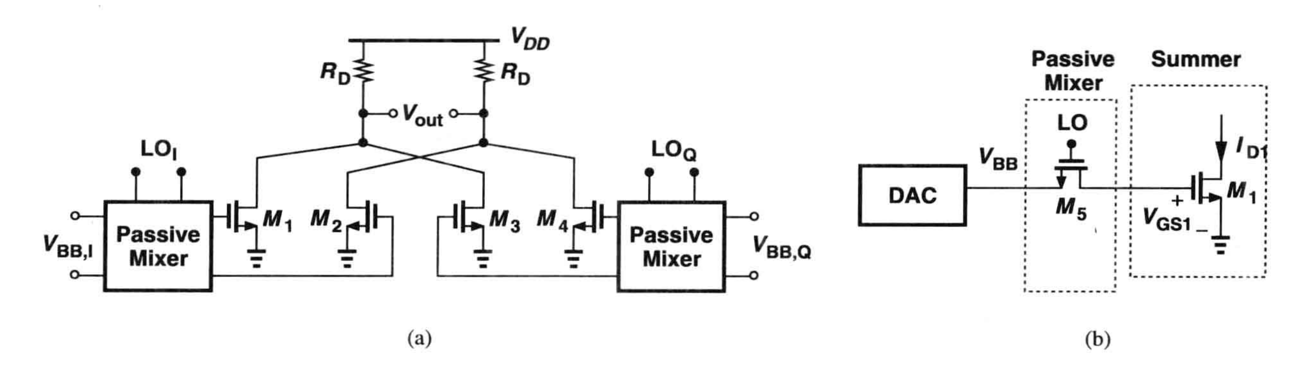


Figure 6.85 *(a) Summation of quadrature outputs, (b) bias definition issue.*

12. The ac ground at the source nodes reduces third order nonlinearity (Chapter 5).

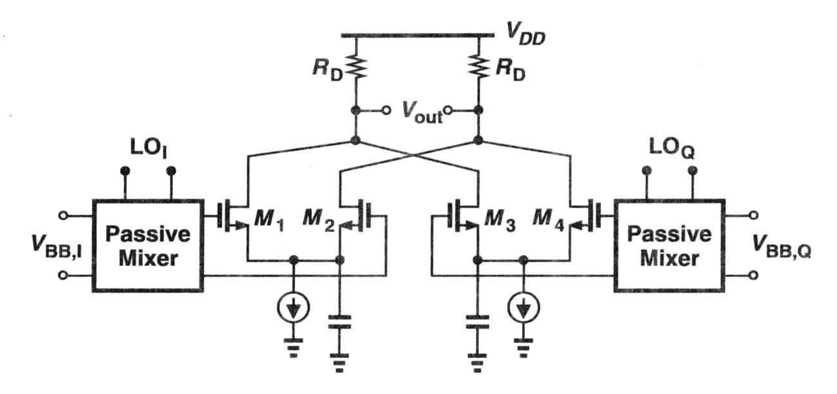


Figure 6.86 *Addition of tail current to define bias of upconversion V/I converters.*

employ ac coupling between the mixer and the V/I converter and define the latter's bias by a current mirror. Alternatively, we can resort to true differential pairs, with their common-source nodes at ac ground (Fig. 6.86). Defined by the tail currents, the bias conditions now remain relatively independent of the input CM level, but each tail current source consumes voltage headroom.

Example 6.30

The trade-off between the voltage drop across R_D in Fig. 6.85(a) and the voltage gain proves undesirable, especially because M_1–M_4 must be biased with some margin with respect to the triode region so as to preserve their linearity in the presence of large signals. Explain how this trade-off can be avoided.

Solution:

Since the output center frequency of the upconverter is typically in the gigahertz range, the resistors can be replaced with inductors. Illustrated in Fig. 6.87, such a technique consumes little headroom (because the dc drop across the inductor is small) and nulls the total capacitance at the output by means of resonance.

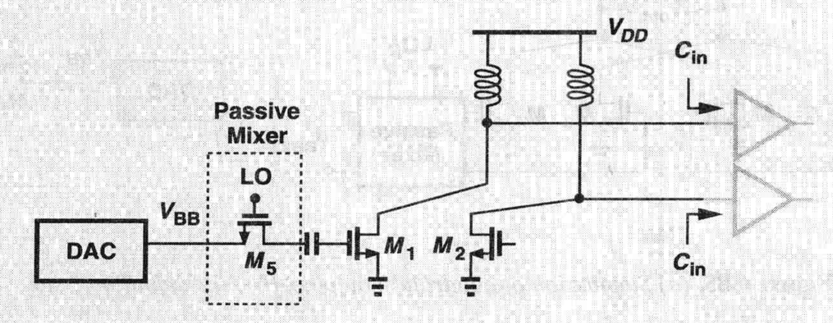


Figure 6.87 *Use of inductive loads to relax upconversion mixer headroom constraints.*

The third issue concerns the available overdrive voltage of the mixer switches, a particularly serious problem in Fig. 6.85(b). We note that M_5 can be ac coupled to M_1, but still requiring a gate voltage of $V_{TH5} + V_{GS1} + V_{BB}$ to turn on. Thus, if the peak LO level is equal to V_{DD}, the switch experiences an overdrive of only $V_{DD} - (V_{TH5} + V_{BB})$, thereby suffering from a tight trade-off between its on-resistance and capacitance. A small overdrive also degrades the linearity of the switch. For example, if $V_{DD} = 1$ V, $V_{TH5} = 0.3$ V, and $V_{BB} = 0.5$ V, then the overdrive is equal to 0.2 V. It is important to recognize that the use of inductors in Fig. 6.87 relaxes the headroom consumption from V_{DD} through R_D and M_1, but the headroom limitation in the path consisting of V_{DD}, V_{GS5}, and V_{BB} still persists.

The foregoing difficulty can be alleviated if the peak LO level can *exceed* V_{DD}. This is accomplished if the LO buffer contains a load inductor tied to V_{DD} (Fig. 6.88).

Now, the dc level of the LO is approximately equal to V_{DD}, with the peak reaching $V_{DD} + V_0$. For example, if $V_{DD} = 1$ V, $V_{TH5} = 0.3$ V, $V_{BB} = 0.5$ V, and $V_0 = 0.5$ V, then the overdrive of M_5 is raised to 0.7 V.

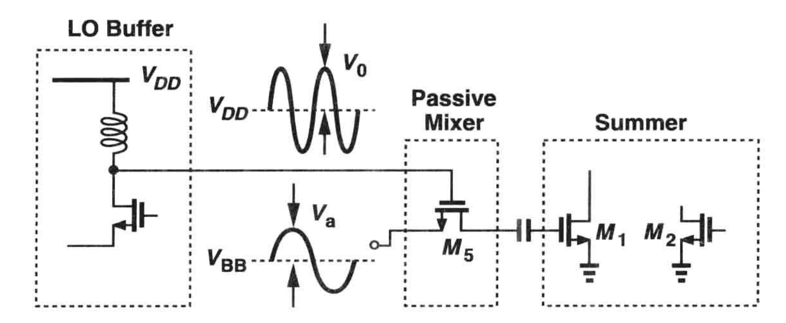


Figure 6.88 *Mixer headroom considerations.*

The above-V_{DD} swings in Fig. 6.88 do raise concern with respect to device voltage stress and reliability. In particular, if the baseband signal has a peak amplitude of V_a and a CM level of V_{BB}, then the gate-source voltage of M_5 reaches a maximum of $V_{DD} + V_0 - (V_{BB} - V_a)$, possibly exceeding the value allowed by the technology. In the above numerical example, since the overdrive of M_5 approaches 0.7 V, $V_{GS5} = 0.7\text{ V} + V_{TH5} = 1$ V in the *absence* of the baseband signal. Thus, if the maximum allowable V_{GS} is 1.2 V, the baseband peak swing is limited to 0.2 V. As explained in Chapter 4, small baseband swings exacerbate the problem of carrier feedthrough in transmitters.

It is important to note that, by now, we have added quite a few inductors to the circuit: one in Fig. 6.84(b) to improve the bandwidth, one in Fig. 6.87 to save voltage headroom, and another in Fig. 6.88 to raise the overdrive of the switches. A quadrature upconverter therefore requires a large number of inductors. The LO buffer in Fig. 6.88 can be omitted if the LO signal is capacitively coupled to the gate of M_5 and biased at V_{DD}.

Carrier Feedthrough It is instructive to study the sources of carrier feedthrough in a transmitter using passive mixers. Consider the baseband interface shown in Fig. 6.89, where the DAC output contains a peak signal swing of V_a and an offset voltage of $V_{OS,DAC}$.

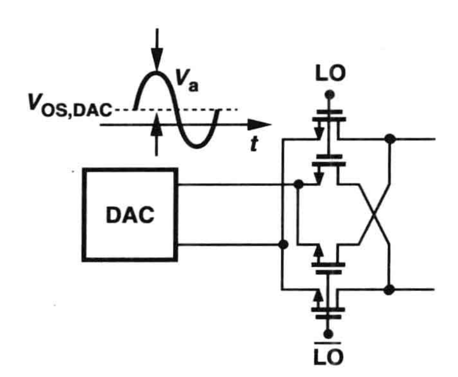


Figure 6.89 *Effect of baseband offset in upconversion mixing.*

An ideal double-balanced passive mixer upconverts both the signal and the offset, producing at its output the RF (or IF) signal and a carrier (LO) component. If modeled as a multiplier, the mixer generates an output given by

$$V_{out}(t) = \alpha(V_a \cos \omega_{BB}t + V_{OS,DAC}) \cos \omega_{LO}t, \tag{6.146}$$

where α is related to the conversion gain. Expanding the right-hand side yields

$$V_{out}(t) = \frac{\alpha V_a}{2} \cos(\omega_{LO} + \omega_{BB})t + \frac{\alpha V_a}{2} \cos(\omega_{LO} - \omega_{BB})t + \alpha V_{OS,DAC} \cos \omega_{LO}t. \tag{6.147}$$

Since $\alpha/2 = 2/\pi$ for a double-balanced mixer, we note that the carrier feedthrough has a peak amplitude of $\alpha V_{OS,DAC} = (4/\pi)V_{OS,DAC}$. Alternatively, we recognize that the relative carrier feedthrough is equal to $\alpha V_{OS,DAC}/(\alpha V_a/2) = 2V_{OS,DAC}/V_a$. For example, if $V_{OS,DAC} = 10\,\text{mV}$ and $V_a = 0.1$ V, then the feedthrough is equal to -34 dB.

Let us now consider the effect of threshold mismatches within the switches themselves. As illustrated in Fig. 6.90(a), the threshold mismatch in one pair shifts the LO waveform vertically, distorting the duty cycle. That is, V_{in}^{+} is multiplied by the equivalent waveforms shown in Fig. 6.90(b). Does this operation generate an output component at f_{LO}? No, carrier feedthrough can occur only if a dc component in the baseband is mixed with the fundamental LO frequency. We therefore conclude that threshold mismatches within passive mixers introduce no carrier feedthrough.[13]

13. The threshold mismatch in fact leads to charge injection mismatch between the switches and a slight disturbance at the output at the LO frequency. But this disturbance carries litter energy because it appears only during LO transitions.

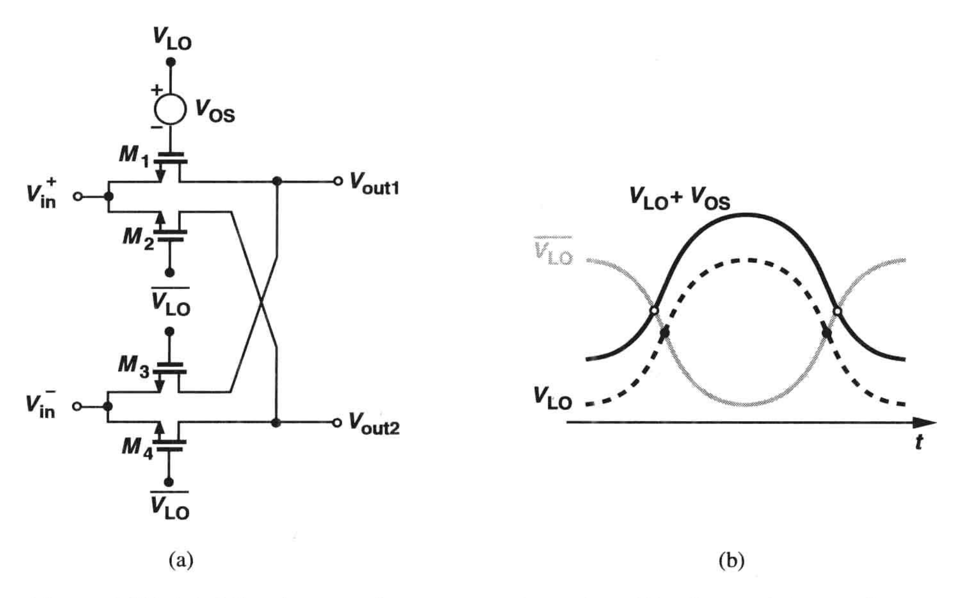


Figure 6.90 *(a) Offset in a passive upconversion mixer, (b) effect on LO waveforms.*

Example 6.31

If asymmetries in the LO circuitry distort the duty cycle, does the passive mixer display carrier feedthrough?

Solution:

In this case, the two switching pairs in Fig. 6.90(a) experience the same duty cycle distortion. The above analysis implies that each pair is free from feedthrough, and hence so does the overall mixer.

The carrier feedthrough in passive upconversion mixers arises primarily from mismatches between the gate-drain capacitances of the switches. As shown in Fig. 6.91, the LO feedthrough observed at X is equal to

$$V_X = V_{LO}\frac{C_{GD1} - C_{GD3}}{C_{GD1} + C_{GD3} + C_X}, \tag{6.148}$$

where C_X denotes the total capacitance seen from X to ground (including the input capacitance of the following stage).

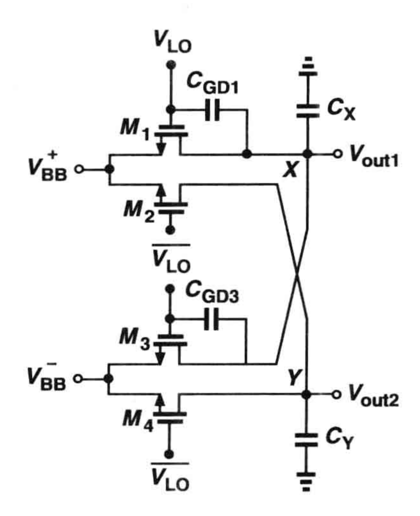


Figure 6.91 *LO feedthrough paths in a passive mixer.*

Example 6.32

Calculate the relative carrier feedthrough for a C_{GD} mismatch of 5%, $C_X \approx 10C_{GD}$, peak LO swing of 0.5 V, and peak baseband swing of 0.1 V.

Solution:

At the output, the LO feedthrough is given by Eq. (6.148) and approximately equal to $(5\%/12)V_{LO} = 2.1$ mV. The upconverted signal has a peak amplitude of $0.1\text{ V} \times (2/\pi) = 63.7$ mV. Thus, the carrier feedthrough is equal to -29.6 dB.

Active Mixers Upconversion in a transmitter can be performed by means of active mixers, facing issues different from those of passive mixers. We begin with a double-balanced topology employing a quasi-differential pair (Fig. 6.92). The inductive loads serve two purposes, namely, they relax voltage headroom issues and raise the conversion gain (and hence the output swings) by nulling the capacitance at the output node. As with active downconversion mixers studied in Section 6.3, the voltage conversion gain can be expressed as

$$A_V = \frac{2}{\pi} g_{m1,2} R_p, \tag{6.149}$$

where R_p is the equivalent parallel resistance of each inductor at resonance.

With only low frequencies present at the gates and drains of M_1 and M_2 in Fig. 6.92, the circuit is quite tolerant of capacitance at nodes P and Q, a point of contrast to downconversion mixers. However, stacking of the transistors limits the voltage headroom. Recall from downconversion mixer calculations in Section 6.3 that the minimum allowable voltage at

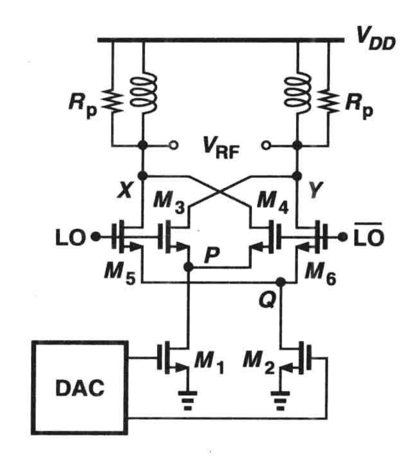


Figure 6.92 *Active upconversion mixer.*

X (or Y) is given by

$$V_{X,min} = V_{GS1} - V_{TH1} + \left(1 + \frac{\sqrt{2}}{2}\right)(V_{GS3} - V_{TH3}), \tag{6.150}$$

if the dc drop across the inductors is neglected. For example, if $V_{GS1} - V_{TH1} = 300\,\text{mV}$ and $V_{GS3} - V_{TH3} = 200\,\text{mV}$, then $V_{X,min} = 640\,\text{mV}$, allowing a peak swing of $V_{DD} - V_{X,min} = 360\,\text{mV}$ at X if $V_{DD} = 1\,\text{V}$. This value is reasonable.

Example 6.33

Equation 6.150 allocates a drain-source voltage to the input transistors equal to their overdrive voltage. Explain why this is inadequate.

Solution:

The voltage gain from each input to the drain of the corresponding transistor is about -1. Thus, as depicted in Fig. 6.93, when one gate voltage rises by V_a, the corresponding drain falls by approximately V_a, driving the transistor into the triode region by $2V_a$. In other words, the V_{DS} of the input devices in the absence of signals must be at least equal to their overdrive voltage plus $2V_a$, further limiting Eq. (6.150) as

$$V_{X,min} = V_{GS1} - V_{TH1} + 2V_a + \left(1 + \frac{\sqrt{2}}{2}\right)(V_{GS3} - V_{TH3}). \tag{6.151}$$

The output swing is therefore small. If $V_a = 100\,\text{mV}$, then the above numerical example yields a peak output swing of 160 mV.

(Continues)

Example 6.33 *(Continued)*

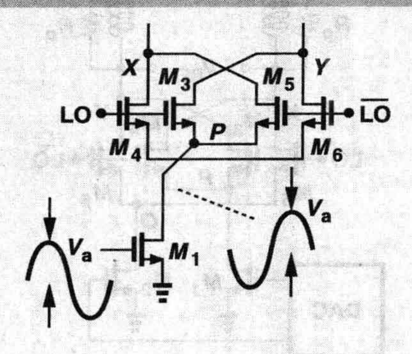


Figure 6.93 *Voltage excursions in an active upconversion mixer.*

Unfortunately, the bias conditions of the circuit of Fig. 6.92 heavily depend on the DAC output common-mode level. Thus, we apply the modification shown in Fig. 6.86, arriving at the topology in Fig. 6.94(a) (a Gilbert cell). This circuit faces two difficulties. First, the current source consumes additional voltage headroom. Second, since node A cannot be held at ac ground by a capacitor at low baseband frequencies, the nonlinearity is more pronounced. We therefore fold the input path and degenerate the differential pair to alleviate these issues [Fig. 6.94(b)].

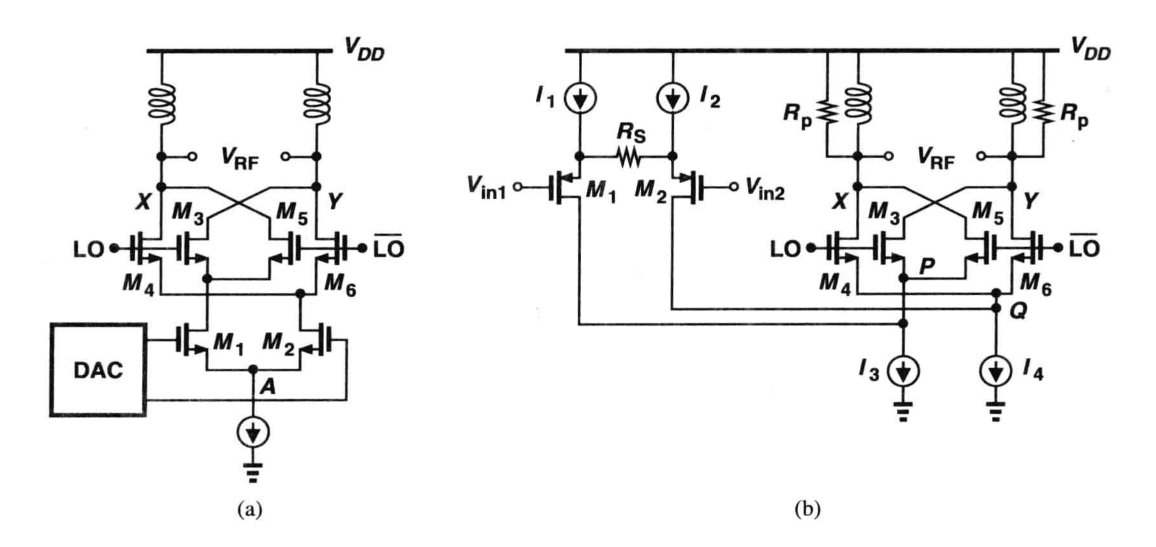


Figure 6.94 *(a) Gilbert cell as upconversion mixer, (b) mixer with folded input stage.*

Example 6.34

Determine the maximum allowable input and output swings in the circuit of Fig. 6.94(b).

Solution:

Let us consider the simplified topology shown in Fig. 6.95. In the absence of signals, the maximum gate voltage of M_1 with respect to ground is equal to $V_{DD}-|V_{GS1}|-|V_{I1}|$, where $|V_{I1}|$ denotes the minimum allowable voltage across I_1. Also, $V_P = V_{I3}$. Note that, due to source degeneration, the voltage gain from the baseband input to P is quite smaller than unity. We therefore neglect the baseband swing at node P. For M_1 to remain in saturation as its gate falls by V_a volts,

$$V_{DD} - |V_{GS1}| - |V_{I1}| - V_a + |V_{TH1}| \geq V_P \tag{6.152}$$

and hence

$$V_a \leq V_{DD} - |V_{GS1} - V_{TH1}| - |V_{I1}| - |V_{I3}|. \tag{6.153}$$

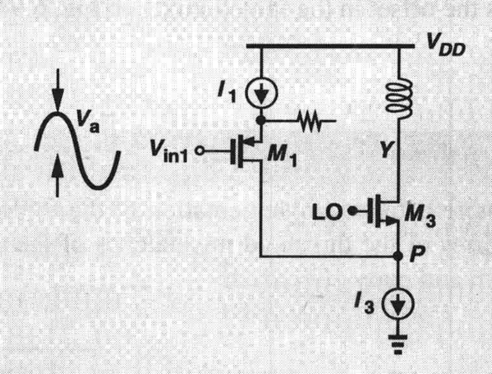


Figure 6.95 *Simplified folded mixer diagram.*

For the output swing, Eq. (6.150) is modified to

$$V_{X,min} = \left(1 + \frac{\sqrt{2}}{2}\right)(V_{GS3} - V_{TH3}) + V_{I3}. \tag{6.154}$$

The tolerable output swing is thus greater than that of the unfolded circuit.

Despite degeneration, the circuit of Fig. 6.94(b) may experience substantial nonlinearity if the baseband voltage swing exceeds a certain value. We recognize that, if $V_{in1} - V_{in2}$ becomes sufficiently negative, $|I_{D1}|$ approaches I_3, starving M_3 and M_5. Now, if the differential input becomes more negative, M_1 and I_1 must enter the triode region so as to satisfy KCL at node P, introducing large nonlinearity. Since the random baseband signal occasionally assumes large voltage excursions, it is difficult to avoid this effect unless the amount

of degeneration (e.g., R_S) is chosen conservatively large, in which case the mixer gain and hence the output swing suffer.

The above observation indicates that the current available to perform upconversion and produce RF swings is approximately equal to the *difference* between I_1 and I_3 (or between I_2 and I_4). The maximum baseband peak single-ended voltage swing is thus given by

$$V_{a,max} = \frac{|I_1 - I_3|}{G_m} \tag{6.155}$$

$$= |I_1 - I_3| \left(\frac{1}{g_{m1,2}} + \frac{R_S}{2} \right). \tag{6.156}$$

Mixer Carrier Feedthrough Transmitters using active upconversion mixers potentially exhibit a higher carrier feedthrough than those incorporating passive topologies. This is because, in addition to the baseband DAC offset, the mixers themselves introduce considerable offset. In the circuits of Figs. 6.92 and 6.94(a), for example, the baseband input transistors suffer from mismatches between their threshold voltages and other parameters. Even more pronounced is the offset in the folded mixer of Fig. 6.94(b), as calculated in the following example.

Example 6.35

Figure 6.96(a) shows a more detailed implementation of the folded mixer. Determine the input-referred offset in terms of the threshold mismatches of the transistor pairs. Neglect channel-length modulation and body effect.

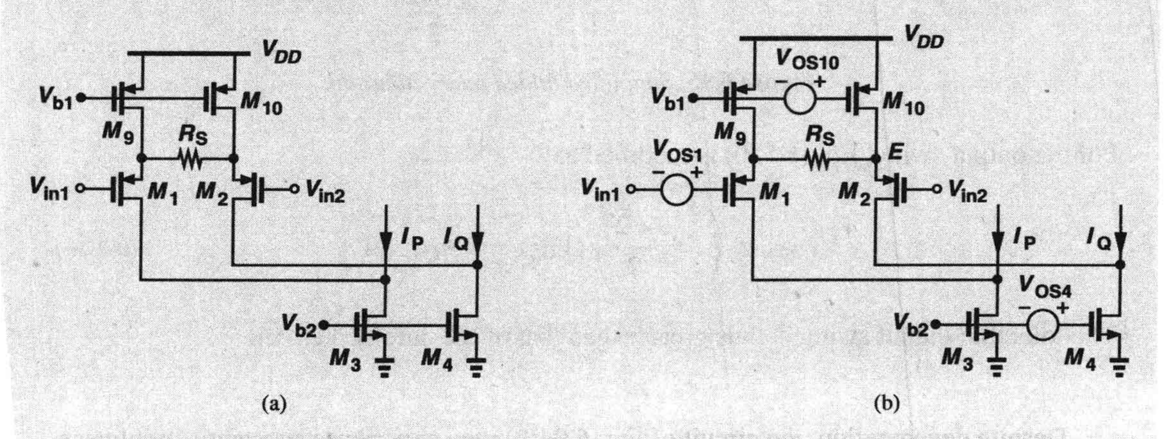


Figure 6.96 *(a) Role of bias current sources in folded mixer, (b) effect of offsets.*

Solution:

As depicted in Fig. 6.96(b), we insert the threshold mismatches and seek the total mismatch between I_P and I_Q. To obtain the effect of V_{OS10}, we first recognize that it generates an additional current of $g_{m10}V_{OS10}$ in M_{10}. This current is split between M_2 and M_1 according

Example 6.35 *(Continued)*

to the small-signal impedance seen at node E, namely,

$$|I_{D2}|_{VOS10} = g_{m10}V_{OS10}\frac{R_S + \dfrac{1}{g_{m1}}}{R_S + \dfrac{1}{g_{m1}} + \dfrac{1}{g_{m2}}} \tag{6.157}$$

$$|I_{D1}|_{VOS10} = g_{m10}V_{OS10}\frac{\dfrac{1}{g_{m2}}}{R_S + \dfrac{1}{g_{m1}} + \dfrac{1}{g_{m2}}}. \tag{6.158}$$

The resulting mismatch between I_P and I_Q is given by the difference between these two:

$$|I_P - I_Q|_{VOS10} = g_{m10}V_{OS10}\frac{R_S}{R_S + \dfrac{2}{g_{m1,2}}}, \tag{6.159}$$

where $g_{m1,2} = g_{m1} = g_{m2}$. Note that this contribution becomes more significant as the degeneration increases, approaching $g_{m10}V_{OS10}$ for $R_S \gg 2/g_{m1,2}$.

The mismatch between M_3 and M_4 simply translates to a current mismatch of $g_{m4}V_{OS4}$. Adding this component to Eq. (6.159), dividing the result by the transconductance of the input pair, $(R_S/2 + 1/g_{m1,2})^{-1}$, and adding V_{OS1}, we arrive at the input-referred offset:

$$V_{OS,in} = g_{m10}R_SV_{OS10} + g_{m4}V_{OS4}\left(\frac{R_S}{2} + \frac{1}{g_{m1,2}}\right) + V_{OS1}. \tag{6.160}$$

This expression imposes a trade-off between the input offset and the overdrive voltages allocated to M_9–M_{10} and M_3–M_4: for a given current, $g_m = 2I_D/(V_{GS} - V_{TH})$ increases as the overdrive decreases, raising $V_{OS,in}$.

In addition to offset, the six transistors in Fig. 6.96(a) also contribute noise, potentially a problem in GSM transmitters.[14] It is interesting to note that LO duty cycle distortion does not cause carrier feedthrough in double-balanced active mixers. This is studied in Problem 6.15.

Active mixers readily lend themselves to quadrature upconversion because their outputs can be summed in the current domain. Figure 6.97 shows an example employing folded mixers.

Design Procedure As mentioned in Section 6.1, the design of upconversion mixers typically follows that of the power amplifier. With the input capacitance of the PA (or PA driver) known, the mixer output inductors, e.g., L_1 and L_2 in Fig. 6.97, are designed to resonate at

14. As explained in Chapter 4, the noise produced by a GSM transmitter in the receive band must be very small.

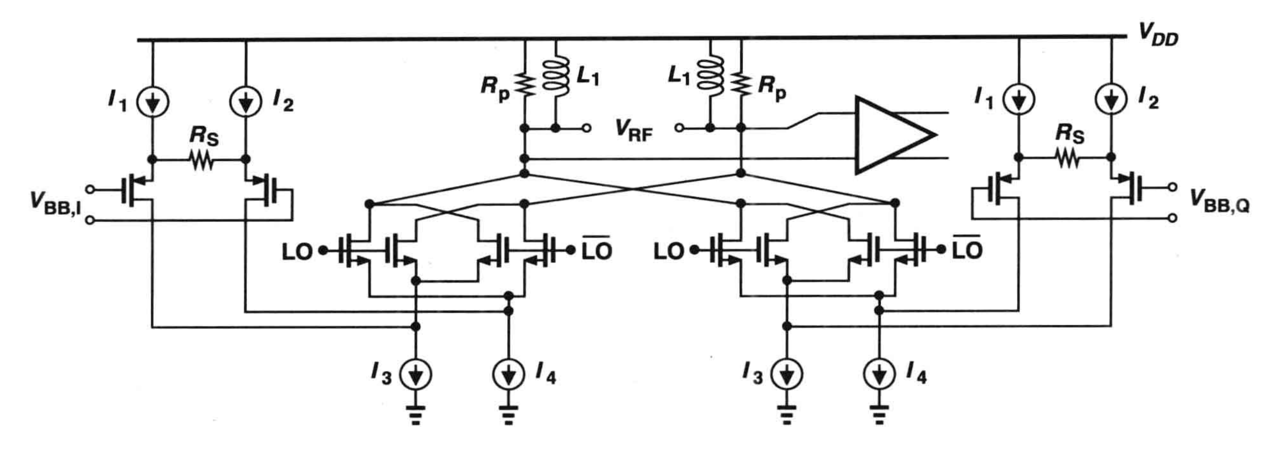


Figure 6.97 *Summation of quadrature outputs.*

the frequency of interest. At this point, the capacitance contributed by the switching quads, C_q, is unknown and must be guessed. Thus,

$$L_1 = L_2 = \frac{1}{\omega_0^2(C_q + C_L)}, \tag{6.161}$$

where C_L includes the input capacitance of the next stage and the parasitic of L_1 or L_2. Also, the finite Q of the inductors introduces a parallel equivalent resistance given by

$$R_p = \frac{Q}{\omega_0(C_q + C_L)}. \tag{6.162}$$

If sensing quadrature baseband inputs with a peak single-ended swing of V_a, the circuit of Fig. 6.97 produces an output swing given by

$$V_{p,out} = \sqrt{2}\frac{2}{\pi}\frac{R_p}{\dfrac{R_S}{2} + \dfrac{1}{g_{mp}}}(2V_a), \tag{6.163}$$

where the factor of $\sqrt{2}$ results from summation of quadrature signals, $2V_a$ denotes the peak differential swing at each input, and g_{mp} is the transconductance of the input PMOS devices. Thus, R_S, g_{mp}, and V_a must be chosen so as to yield both the required output swing and proper linearity.

How do we choose the bias currents? We must first consider the following example.

Example 6.36

The tail current of Fig. 6.98 varies with time as $I_{SS} = I_0 + I_0 \cos \omega_{BB} t$. Calculate the voltage swing of the upconverted signal.

Example 6.36 *(Continued)*

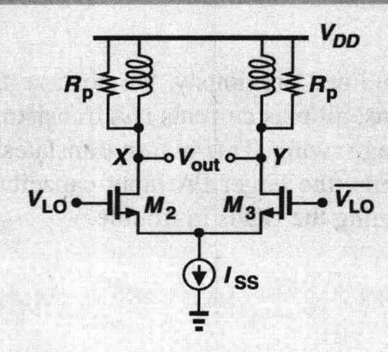


Figure 6.98 *Simplified stage for swing calculation.*

Solution:

We know that I_{SS} is multiplied by $(2/\pi)R_p$ as it is upconverted. Thus, the output voltage swing at $\omega_{LO} - \omega_{BB}$ or $\omega_{LO} + \omega_{BB}$ is equal to $(2/\pi)I_0R_p$. We have assumed that I_{SS} swings between zero and $2I_0$, but an input transistor experiencing such a large current variation may become quite nonlinear.

The above example suggests that I_0 must be sufficiently large to yield the required output swing. That is, with R_p known, I_0 can be calculated. A double-balanced version of the circuit generates twice the output swing, and a quadrature topology (Fig. 6.97) raises the result by another factor of $\sqrt{2}$, delivering a peak output swing of $(4\sqrt{2}/\pi)I_0R_P$. With I_0 ($= I_3/2 = I_4/2$ in Fig. 6.97) known, we select $I_1 = I_2 = I_3/2 = I_4/2$.

How do we select the transistor dimensions? Let us first consider the switching devices, noting that each switching pair in Fig. 6.97 carries a current of nearly I_3 ($= I_4$) at the extremes of the baseband swings. These transistors must therefore be chosen wide enough to (1) carry a current of I_3 while leaving adequate voltage headroom for I_3 and I_4, and (2) switch their tail currents nearly completely with a given LO swing.

Next, the transistors implementing I_3 and I_4 are sized according to their allowable voltage headroom. Lastly, the dimensions of the input differential pair and the transistors realizing I_1 and I_2 are chosen. With these choices, the input-referred offset [Eq. (6.160)] must be checked.

Example 6.37

An engineer designs a quadrature upconversion mixer for a given output frequency, a given output swing, and a given load capacitance, C_L. Much to her dismay, the engineer's manager raises C_L to $2C_L$ because the following power amplifier must be redesigned for a higher output power. If the upconverter output swing must remain the same, how can the engineer modify her design to drive $2C_L$?

(Continues)

Example 6.37 *(Continued)*

Solution:

Following the calculations outlined previously, we observe that the load inductance and hence R_p must be halved. Thus, all bias currents and transistor widths must be doubled so as to maintain the output voltage swing. This in turn translates to a higher load capacitance seen by the LO. In other words, the larger PA input capacitance "propagates" to the LO port. Now, the engineer designing the LO is in trouble.

REFERENCES

[1] B. Razavi, "A Millimeter-Wave Circuit Technique," *IEEE J. of Solid-State Circuits,* vol. 43, pp. 2090–2098, Sept. 2008.

[2] P. Eriksson and H. Tenhunen, "The Noise Figure of A Sampling Mixer: Theory and Measurement," *IEEE Int. Conf. Electronics, Circuits, and Systems,* pp. 899–902, Sept. 1999.

[3] S. Zhou and M. C. F. Chang, "A CMOS Passive Mixer with Low Flicker Noise for Low-Power Direct-Conversion Receivers," *IEEE J. of Solid-State Circuits,* vol. 40, pp. 1084, 1093, May 2005.

[4] D. Leenaerts and W. Readman-White, "1/f Noise in Passive CMOS Mixers for Low and Zero IF Integrated Receivers," *Proc. ESSCIRC,* pp. 41–44, Sept. 2001.

[5] A. Mirzaei et al., "Analysis and Optimization of Current-Driven Passive Mixers in Narrow-band Direct-Conversion Receivers," *IEEE J. of Solid-State Circuits,* vol. 44, pp. 2678–2688, Oct. 2009.

[6] D. Kaczman et al., "A Single-Chip 10-Band WCDMA/HSDPA 4-Band GSM/EDGE SAW-less CMOS Receiver with DigRF 3G Interface and +90-dBm IIP2," *IEEE J. Solid-State Circuits,* vol. 44, pp. 718–739, March 2009.

[7] H. Darabi and A. A. Abidi, "Noise in RF-CMOS Mixers: A Simple Physical Model," *IEEE J. of Solid-State Circuits,* vol. 35, pp. 15–25, Jan. 2000.

[8] W. H. Sansen and R. G. Meyer, "Distortion in Bipolar Transistor Variable-Gain Amplifiers," *IEEE Journal of Solid-State Circuits,* vol. 8, pp. 275–282, Aug. 1973.

[9] B. Razavi, "A 60-GHz CMOS Receiver Front-End," *IEEE J. of Solid-State Circuits,* vol. 41, pp. 17–22, Jan. 2006.

[10] B. Razavi, "A 900-MHz CMOS Direct-Conversion Receiver," *Dig. of Symposium on VLSI Circuits,* pp. 113–114, June 1997.

[11] M. Brandolini et al., "A +78-dBm IIP2 CMOS Direct Downconversion Mixer for Fully-Integrated UMTS Receivers," *IEEE J. Solid-State Circuits,* vol. 41, pp. 552–559, March 2006.

[12] D. Manstretta, M. Brandolini, and F. Svelto, "Second-Order Intermodulation Mechanisms in CMOS Downconverters," *IEEE J. Solid-State Circuits,* vol. 38, pp. 394–406, March 2003.

[13] H. Darabi and J. Chiu, "A Noise Cancellation Technique in Active RF-CMOS Mixers," *IEEE J. of Solid-State Circuits,* vol. 40, pp. 2628–2632, Dec. 2005.

[14] R. S. Pullela, T. Sowlati, and D. Rozenblit, "Low Flicker Noise Quadrature Mixer Topology," *ISSCC Dig. Tech. Papers,* pp. 76–77, Feb. 2006.

[15] B. Razavi, "CMOS Transceivers for the 60-GHz Band," *IEEE Radio Frequency Integrated Circuits Symposium,* pp. 231–234, June 2006.

PROBLEMS

6.1. Suppose in Fig. 6.13, the LNA has a voltage gain of A_0 and the mixers have a high input impedance. If the I and Q outputs are simply added, determine the overall noise figure in terms of the NF of the LNA and the input-referred noise voltage of the mixers.

6.2. Making the same assumptions as in the above problem, determine the noise figure of a Hartley receiver. Neglect the noise of the 90°-phase-shift circuit and the output adder.

6.3. Consider the circuit of Fig. 6.99, where C_1 and C_2 are identical and represent the gate-source capacitances in Fig. 6.15(b). Assume $V_1 = -V_2 = V_0 \cos \omega_{LO} t$.

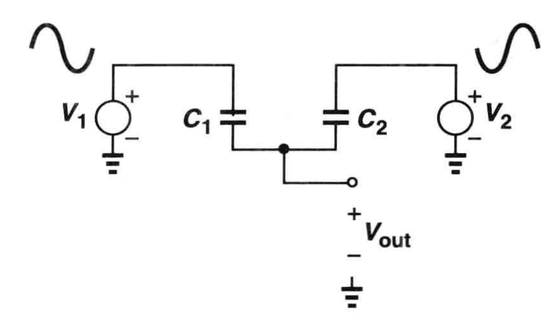


Figure 6.99 *Capacitors driven by differential waveforms.*

(a) If $C_1 = C_2 = C_0(1 + \alpha_1 V)$, where V denotes the voltage across each capacitor, determine the LO feedthrough component(s) in V_{out}. Assume $\alpha_1 V \ll 1$.

(b) Repeat part (a) if $C_1 = C_2 = C_0(1 + \alpha_1 V + \alpha_2 V^2)$.

6.4. We express V_{n1} in Fig. 6.29(c) as the product of the shaped resistor noise *voltage* and a square wave toggling between 0 and 1. Prove that the spectrum of V_{n1} is given by Eq. (6.31).

6.5. Prove that the voltage conversion gain of a sampling mixer approaches 6 dB as the width of the LO pulses tends to zero (i.e., as the hold time approaches the LO period).

6.6. Consider the LO buffer shown in Fig. 6.55. Prove that the noise of M_5 and M_6 appears differentially at nodes A and B (but the noise due to the loss of the tanks does not).

6.7. In the active mixer of Fig. 6.57, $I_{n,M1}$ contains all frequency components. Prove that the convolution of these components with the harmonics of the LO in essence multiplies $4kT\gamma/g_m$ by a factor of $\pi^2/4$.

6.8. If transistors M_2 and $M3$ in Fig. 6.60(a) have a threshold mismatch of V_{OS}, determine the output flicker noise due to the flicker noise of I_{SS}.

6.9. Shown in Fig. 6.100 is the front end of a 1.8-GHz receiver. The LO frequency is chosen to be 900 MHz and the load inductors and capacitances resonate with a quality

factor of Q at the IF. Assume M_1 is biased at a current of I_1, and the mixer and the LO are perfectly symmetric.

(a) Assuming M_2 and M_3 switch abruptly and completely, compute the LO-IF feedthrough, i.e., the measured level of the 900-MHz output component in the absence of an RF signal.

(b) Explain why the flicker noise of M_1 is critical here.

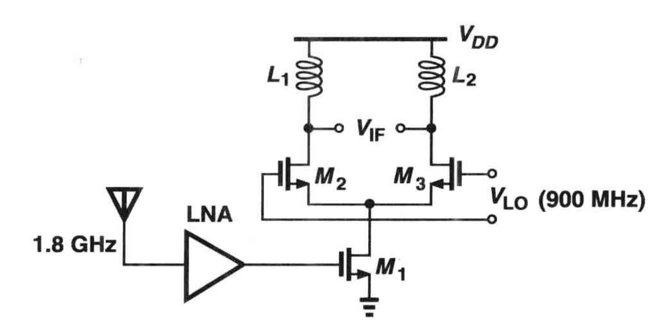


Figure 6.100 *Front-end chain for a 1.8-GHz RX.*

6.10. Suppose the helper in Fig. 6.67 reduces the bias current of the switching pair by a factor of 2. By what factor does the input-referred contribution of the flicker noise fall?

6.11. In the circuit of Fig. 6.67, we place a parallel RLC tank in series with the source of M_4 such that, at resonance, the noise contribution of M_4 is reduced. Recalculate Eq. (6.116) if the tank provides an equivalent parallel resistance of R_p. (Bear in mind that R_p itself produces noise.)

6.12. Can the circuit of Fig. 6.81(a) be viewed as a differential pair whose tail current is modulated at a rate of $2f_{LO}$? Carry out the analysis and explain your result.

6.13. Suppose the quadrature upconversion mixers in a GSM transmitter operate with a peak baseband swing of 0.3 V. If the TX delivers an output power of 1 W, determine the maximum tolerable input-referred noise of the mixers such that the transmitted noise in the GSM RX band does not exceed -155 dBm.

6.14. Prove that the voltage conversion gain of a single-balanced return-to-zero mixer is equal to $2/\pi$ even for upconversion.

6.15. Prove that LO duty cycle distortion does not introduce carrier feedthrough in double-balanced active mixers.

6.16. The circuit shown in Fig. 6.101 is a dual-gate mixer used in traditional microwave design. Assume when M_1 is on, it has an on-resistance of R_{on1}. Also, assume abrupt edges and a 50% duty cycle for the LO and neglect channel-length modulation and body effect.

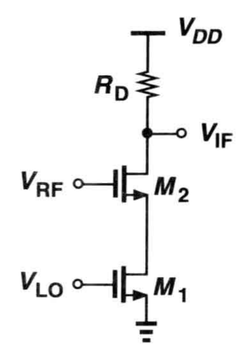


Figure 6.101 *Dual-gate mixer.*

(a) Compute the voltage conversion gain of the circuit. Assume M_2 does not enter the triode region and denote its transconductance by g_{m2}.

(b) If R_{on1} is very small, determine the IP_2 of the circuit. Assume M_2 has an overdrive of $V_{GS0} - V_{TH}$ in the absence of signals (when it is on).

6.17. Consider the active mixer shown in Fig. 6.102, where the LO has abrupt edges and a 50% duty cycle. Also, channel-length modulation and body effect are negligible. The load resistors exhibit mismatch, but the circuit is otherwise symmetric. Assume M_1 carries a bias current of I_{SS}.

(a) Determine the output offset voltage.

(b) Determine the IP_2 of the circuit in terms of the overdrive and bias current of M_1.

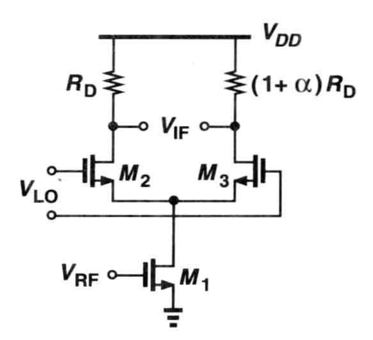
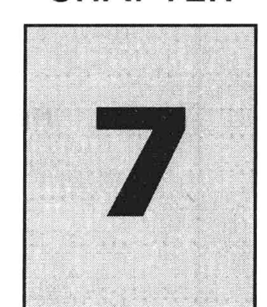


Figure 6.102 *Active mixer with load mismatch.*

CHAPTER



PASSIVE DEVICES

An important factor in the success of today's RF integrated circuits has been the ability to incorporate numerous on-chip passive devices, thus reducing the number of off-chip components. Of course, some integrated passive devices, especially in CMOS technology, exhibit a lower quality than their external counterparts. But, as seen throughout this book, we now routinely use hundreds of such devices in RF transceiver design—an impractical paradigm if they were placed off-chip.

This chapter deals with the analysis and design of integrated inductors, transformers, varactors, and constant capacitors. The outline of the chapter is shown below.

Inductors	Inductor Structures	Transformers	Varactors
Basic Structure	Symmetric Inductors	Structures	PN Junctions
Inductance Equations	Effect of Ground Shield	Effect of Coupling Capacitance	MOS Varactors
Parasitic Capacitances	Stacked Spirals	Transformer Modeling	Varactor Modeling
Loss Mechanisms			
Inductor Modeling			

7.1 GENERAL CONSIDERATIONS

While analog integrated circuits commonly employ resistors and capacitors, RF design demands additional passive devices, e.g., inductors, transformers, transmission lines, and varactors. Why do we insist on integrating these devices on the chip? If the entire transceiver requires only one or two inductors, why not utilize bond wires or external components? Let us ponder these questions carefully.

Modern RF design needs *many* inductors. To understand this point, consider the simple common-source stage shown in Fig. 7.1(a). This topology suffers from two serious drawbacks: (a) the bandwidth at node X is limited to $1/[(R_D||r_{O1})C_D]$, and (b) the voltage headroom trades with the voltage gain, $g_{m1}(R_D||r_{O1})$. CMOS technology scaling tends to improve the former but at the cost of the latter. For example, in 65-nm technology with a

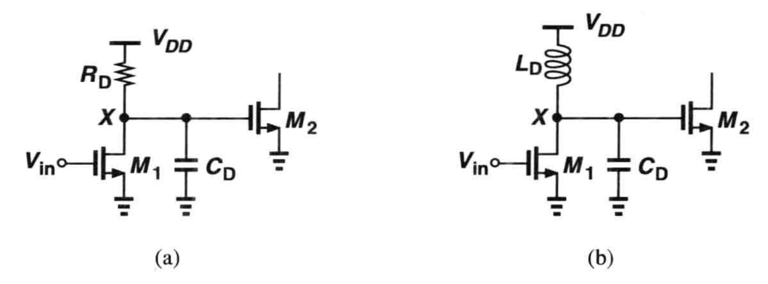


Figure 7.1 *CS stage with (a) resistive, and (b) inductive loads.*

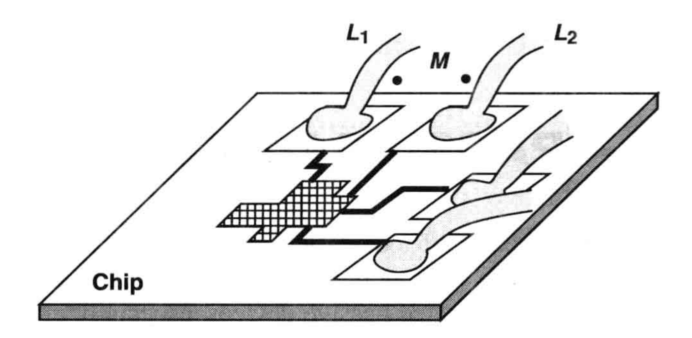


Figure 7.2 *Coupling between bond wires.*

1-V supply, the circuit provides a bandwidth of several gigahertz but a voltage gain in the range of 3 to 4.

Now consider the inductively-loaded stage depicted in Fig. 7.1(b). Here, L_D resonates with C_D, allowing operation at much higher frequencies (albeit in a narrow band). Moreover, since L_D sustains little dc voltage drop, the circuit can comfortably operate with low supply voltages while providing a reasonable voltage gain (e.g., 10). Owing to these two key properties, inductors have become popular in RF transceivers. In fact, the ability to integrate inductors has encouraged RF designers to utilize them almost as extensively as other devices such as resistors and capacitors.

In addition to cost penalties, the use of off-chip devices entails other complications. First, the bond wires and package pins connecting the chip to the outside world may experience significant coupling (Fig. 7.2), creating crosstalk between different parts of the transceiver.

Example 7.1

Identify two undesirable coupling mechanisms if the LO inductor is placed off-chip.

Solution:

As illustrated in Fig. 7.3, the bond wire leading to the inductor couples to the LNA input bond wire, producing LO emission and large dc offsets in the baseband. Additionally, the coupling from the PA output bond wire may result in severe LO pulling.

Example 7.1 *(Continued)*

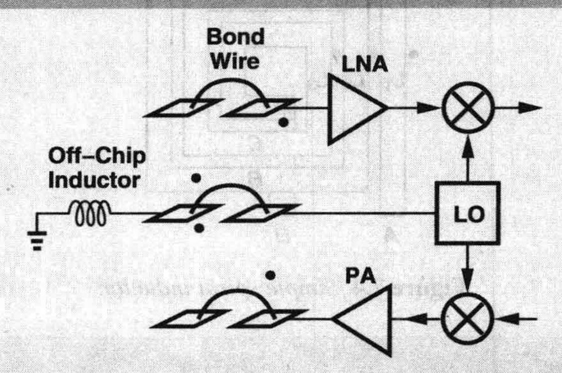


Figure 7.3 *Hypothetical transceiver using an off-chip inductor.*

Second, external connections introduce parasitics that become significant at higher frequencies. For example, a 1-nH bond wire inductance considerably alters the behavior of gigahertz circuits. Third, it is difficult to realize differential operation with external loads because of the poor control of the length of bond wires.

Despite the benefits of integrated components, a critical challenge in RF microelectronics has been how to design high-performance circuits with relatively *poor* passive devices. For example, on-chip inductors exhibit a lower quality factor than their off-chip counterparts, leading to higher "phase noise" in oscillators (Chapter 8). The RF designer must therefore seek new oscillator topologies that produce a low phase noise even with a moderate inductor Q.

Modeling Issues Unlike integrated resistors and parallel-plate capacitors, which can be characterized by a few simple parameters, inductors and some other structures are much more difficult to model. In fact, the required modeling effort proves a high barrier to entry into RF design: one cannot add an inductor to a circuit without an accurate model, and the model heavily depends on the geometry, the layout, and the technology's metal layers (which is the thickest).

It is for these considerations that we devote this chapter to the analysis and design of passive devices.

7.2 INDUCTORS

7.2.1 Basic Structure

Integrated inductors are typically realized as metal spirals (Fig. 7.4). Owing to the mutual coupling between every two turns, spirals exhibit a higher inductance than a straight line having the same length. To minimize the series resistance and the parasitic capacitance, the spiral is implemented in the top metal layer (which is the thickest).

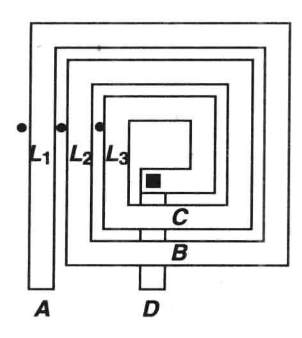


Figure 7.4 *Simple spiral inductor.*

Example 7.2

For the three-turn spiral shown in Fig. 7.4, determine the overall inductance.

Solution:

We identify the three turns as *AB*, *BC*, and *CD*, denoting their individual inductances by L_1, L_2, and L_3, respectively. Also, we represent the mutual inductance between L_1 and L_2 by M_{12}, etc. Thus, the total inductance is given by

$$L_{tot} = L_1 + L_2 + L_3 + M_{12} + M_{13} + M_{23}. \tag{7.1}$$

Equation (7.1) suggests that the total inductance rises in proportion to the *square* of the number of turns. In fact, we prove in Problem 7.1 that the inductance expression for an N-turn structure contains $N(N+1)/2$ terms. However, two factors limit the growth rate as a function of N: (a) due to the geometry's planar nature, the inner turns are smaller and hence exhibit lower inductances, and (b) the mutual coupling factor is only about 0.7 for adjacent turns, falling further for non-adjacent turns. For example, in Eq. (7.1), L_3 is quite smaller than L_1, and M_{13} quite smaller than M_{12}. We elaborate on these points in Example 7.4.

A two-dimensional square spiral is fully specified by four quantities (Fig. 7.5): the outer dimension, D_{out}, the line width, W, the line spacing, S, and the number of turns, N.[1]

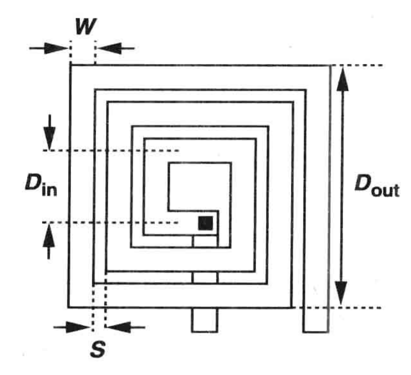


Figure 7.5 *Various dimensions of a spiral inductor.*

1. One may use the inner opening dimension, D_{in}, rather than D_{out} or N.

The inductance primarily depends on the number of turns and the diameter of each turn, but the line width and spacing indirectly affect these two parameters.

Example 7.3

The line width of a spiral is doubled to reduce its resistance; D_{out}, S, and N remain constant. How does the inductance change?

Solution:

As illustrated in Fig. 7.6, the doubling of the width inevitably decreases the diameter of the inner turns, thus lowering their inductance, and the larger spacing between the legs reduces their mutual coupling. We note that further increase in W may also lead to *fewer* turns, reducing the inductance.

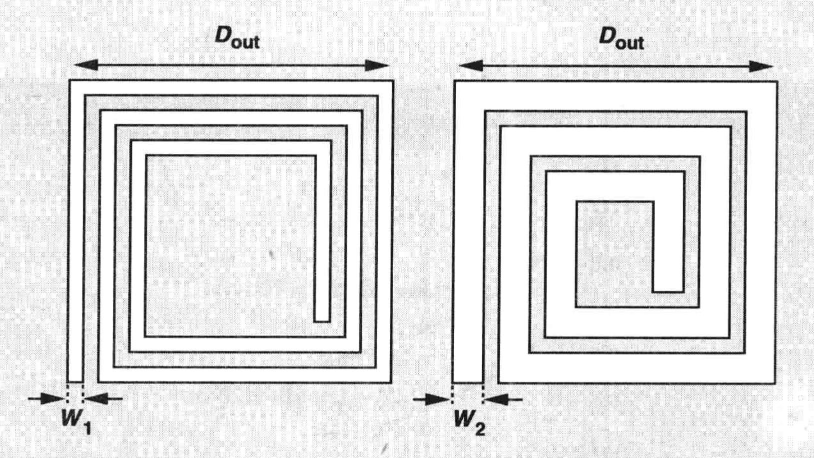


Figure 7.6 *Effect of doubling line width of a spiral.*

Compared with transistors and resistors, inductors typically have much greater dimensions ("foot prints"), resulting in a large chip area and long interconnects traveling from one block to another. It is therefore desirable to minimize the outer dimensions of inductors. For a given inductance, this can be accomplished by (a) decreasing W [Fig. 7.7(a)], or (b) increasing N [Fig. 7.7(b)]. In the former case, the line resistance rises, degrading the inductor quality. In the latter case, the mutual coupling between the sides of the innermost turns *reduces* the inductance because opposite sides carry currents in opposite directions. As shown in Fig. 7.7(b), the two opposite legs of the innermost turn produce opposing magnetic fields, partially cancelling each other's inductance.

Example 7.4

Figure 7.8 plots the magnetic coupling factor between two straight metal lines as a function of their normalized spacing, S/W. Obtained from electromagnetic field simulations, the plots correspond to two cases: each line is 20 μm or 100 μm long. (The line width is 4 μm.) What inner diameter do these plots prescribe for spiral inductors?

(Continues)

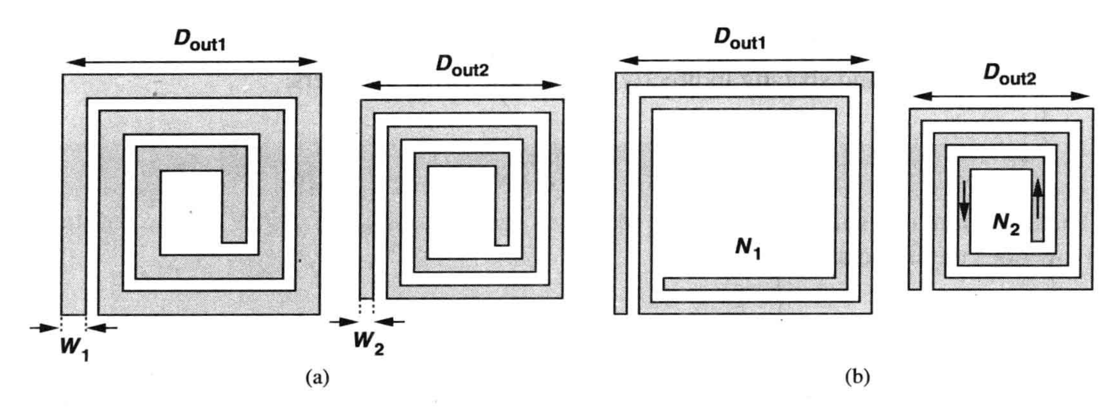


Figure 7.7 *Effect of (a) reducing the outer dimension and the line width, or (b) reducing the outer dimension and increasing the number of turns.*

Example 7.4 (*Continued*)

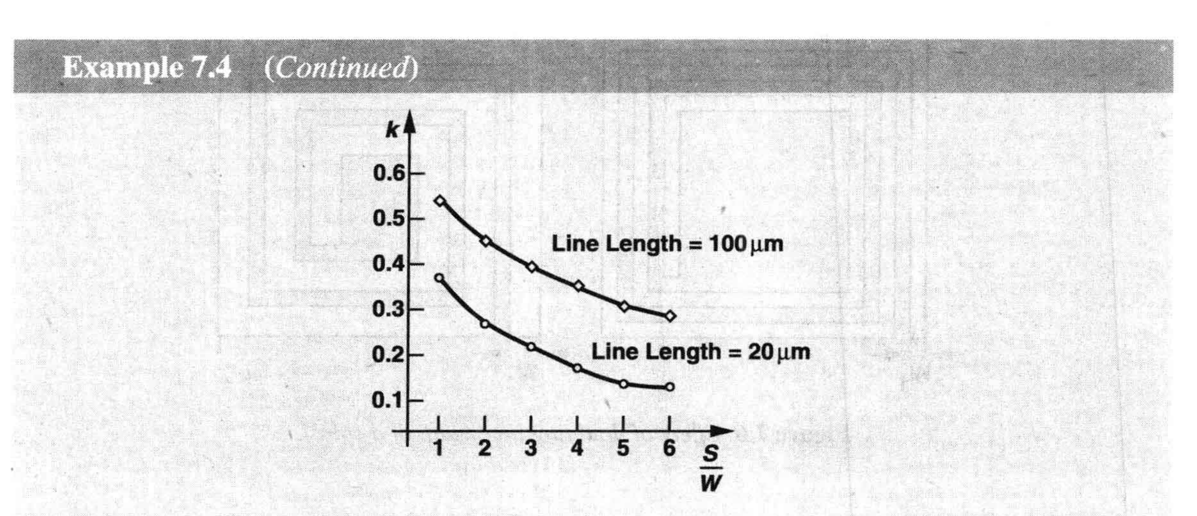


Figure 7.8 *Coupling factor between two straight lines as a function of their normalized spacing.*

Solution:

We wish to minimize the coupling between the opposite sides of the innermost turn. Relevant to typical inductor designs is the plot for a line length of 20 μm, suggesting that a diameter of 5 to 6 times W should be chosen for the inner opening to ensure negligible coupling. It is helpful to remember this rule of thumb.

Even for the basic inductor structure of Fig. 7.5, we must answer a number of questions: (1) How are the inductance, the quality factor, and the parasitic capacitance of the structure calculated? (2) What trade-offs do we face in the choice of these values? (3) What technology and inductor parameters affect the quality factor? These questions are answered in the context of inductor modeling in Section 7.2.6.

7.2.2 Inductor Geometries

Our qualitative study of the square spiral inductors reveals some degrees of freedom in the design, particularly the number of turns and the outer dimension. But there are many other inductor geometries that further add to the design space.

Figure 7.9 shows a collection of inductor structures encountered in RF IC design. We investigate the properties of these topologies later in this chapter, but the reader can observe at this point that: (1) the structures in Figs. 7.9(a) and (b) depart from the square shape, (2) the spiral in Fig. 7.9(c) is *symmetric*, (3) the "stacked" geometry in Fig. 7.9(d) employs two or more spirals in *series*, (4) the topology in Fig. 7.9(e) incorporates a grounded "shield" under the inductor, and (5) the structure in Fig. 7.9(f) places two or more spirals in *parallel*.[2] Of course, many of these concepts can be combined, e.g., the parallel topology of Fig. 7.9(f) can also utilize symmetric spirals and a grounded shield.

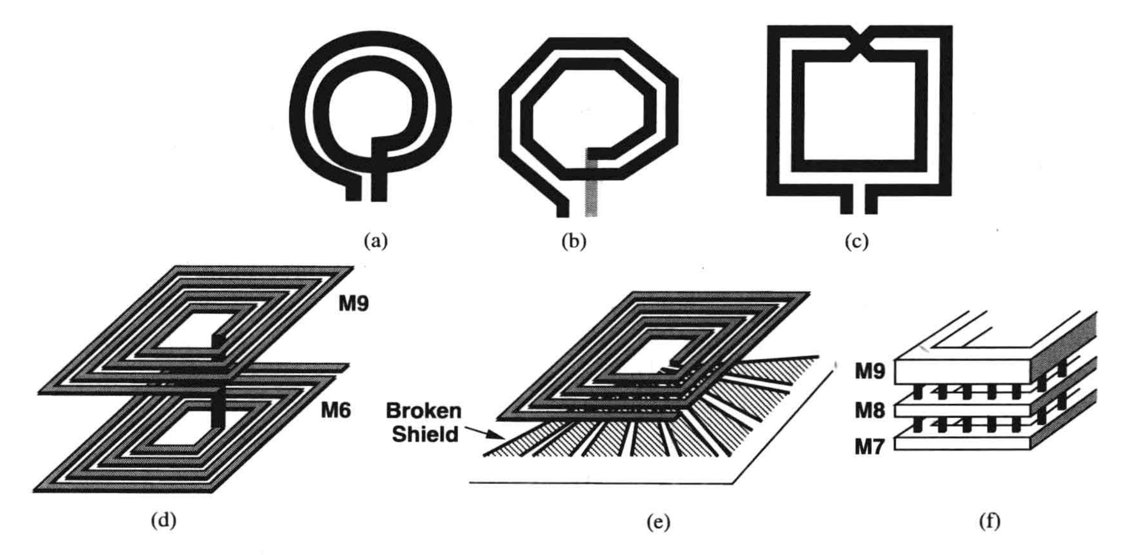


Figure 7.9 *Various inductor structures: (a) circular, (b) octagonal, (c) symmetric, (d) stacked, (e) with grounded shield, (f) parallel spirals.*

Why are there so many different inductor structures? These topologies have resulted from the vast effort expended on improving the trade-offs in inductor design, specifically those between the quality factor and the capacitance or between the inductance and the dimensions.

While providing additional degrees of freedom, the abundance of the inductor geometries also complicates the modeling task, especially if laboratory *measurements* are necessary to fine-tune the theoretical models. How many types of inductors and how many different values must be studied? Which structures are more promising for a given circuit application? Facing practical time limits, designers often select only a few geometries and optimize them for their circuit and frequency of interest.

2. The spirals are shorted to one another by vias, although the vias are not necessary.

7.2.3 Inductance Equations

With numerous inductors used in a typical transceiver, it is desirable to have closed-form equations that provide the inductance value in terms of the spiral's geometric properties. Indeed, various inductance expressions have been reported in the literature [1–3], some based on curve fitting and some based on physical properties of inductors. For example, an empirical formula that has less than 10% error for inductors in the range of 5 to 50 nH is given in [1] and can be reduced to the following form for a square spiral:

$$L \approx 1.3 \times 10^{-7} \frac{A_m^{5/3}}{A_{tot}^{1/6} W^{1.75} (W+S)^{0.25}}, \tag{7.2}$$

where A_m is the metal area (the shaded area in Fig. 7.5) and A_{tot} is the total inductor area ($\approx D_{out}^2$ in Fig. 7.5). All units are metric.

Example 7.5

Calculate the inductor metal area in terms of the other geometric properties.

Solution:

Consider the structure shown in Fig. 7.10. We say this spiral has three turns because each of the four sides contains three complete legs. To determine the metal area, we compute the total length, l_{tot}, of the wire and multiply it by W. The length from A to B is equal to D_{out}, from B to C, equal to $D_{out} - W$, etc. That is,

$$l_{AB} = D_{out} \tag{7.3}$$

$$l_{BC} = l_{CD} = D_{out} - W \tag{7.4}$$

$$l_{DE} = l_{EF} = D_{out} - (2W + S) \tag{7.5}$$

$$l_{FG} = l_{GH} = D_{out} - (3W + 2S) \tag{7.6}$$

$$l_{HI} = l_{IJ} = D_{out} - (4W + 3S) \tag{7.7}$$

$$l_{JK} = l_{KL} = D_{out} - (5W + 4S) \tag{7.8}$$

$$l_{LM} = D_{out} - (6W + 5S). \tag{7.9}$$

Adding these lengths and generalizing the result for N turns, we have

$$l_{tot} = 4ND_{out} - 2W[1 + 2 + \cdots + (2N-1)] - 2NW$$

$$- 2S[1 + 2 + \cdots + (2N-2)] - (2N-1)S \tag{7.10}$$

$$= 4ND_{out} - 4N^2W - (2N-1)^2 S. \tag{7.11}$$

Since $l_{tot} \gg S$, we can add one S to the right-hand side so as to simplify the expression:

$$l_{tot} \approx 4N[D_{out} - W - (N-1)(W+S)]. \tag{7.12}$$

Example 7.5 *(Continued)*

The metal area is thus given by

$$A_m = W[4ND_{out} - 4N^2W - (2N-1)^2S] \tag{7.13}$$

$$\approx 4NW[D_{out} - W - (N-1)(W+S)]. \tag{7.14}$$

This equation is also used for calculating the area capacitance of the spiral.

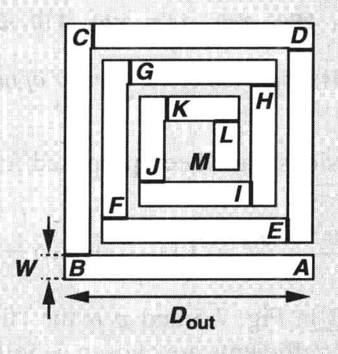


Figure 7.10 *Spiral inductor for calculation of line length.*

An interesting property of inductors is that, for a given wire length, width, and spacing, their inductance is a weak function of the number of turns. This can be seen by finding D_{out} from (7.12), noting that $A_{tot} \approx D_{out}^2$, and manipulating (7.2) as follows:

$$L \approx 1.3 \times 10^{-7} \frac{l_{tot}^{5/3}}{\left[\frac{l_{tot}}{4N} + W + (N-1)(W+S)\right]^{1/3} W^{0.083}(W+S)^{0.25}}. \tag{7.15}$$

We observe that N appears only within the square brackets in the denominator, in two terms varying in opposite directions, with the result raised to the power of $1/3$. For example, if $l_{tot} = 2000\,\mu\text{m}$, $W = 4\,\mu\text{m}$, and $S = 0.5\,\mu\text{m}$, then as N varies from 2 to 3 to 4 to 5, then inductance rises from 3.96 nH to 4.47 nH to 4.83 nH to 4.96 nH, respectively. In other words, a given length of wire yields roughly a constant inductance regardless of how it is "wound."[3] The key point here is that, since this length has a given series resistance (at low frequencies), the choice of N only mildly affects the Q (but can save area).

Figure 7.11 plots the inductance predicted by the simulator ASITIC (described below) as N varies from 2 to 6 and the total wire length remains at $2000\,\mu\text{m}$.[4] We observe that L becomes relatively constant for $N > 3$. Also, the values produced by ASITIC are lower than those given by Eq. (7.15).

3. But the number of turns must be at least 2 to create mutual coupling.
4. The outer dimension varies from 260 μm to 110 μm in this experiment.

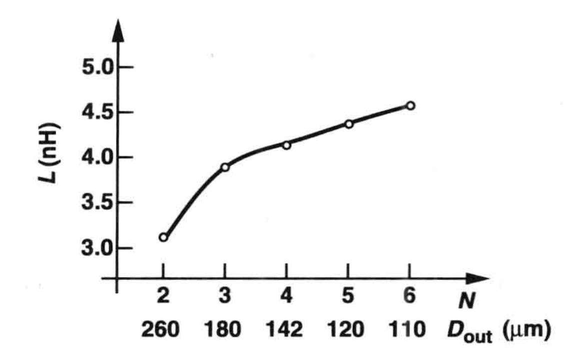


Figure 7.11 *Inductance as a function of the number of turns for a given line length.*

A number of other expressions have been proposed for the inductance of spirals. For example,

$$L = \frac{\mu_0 N^2 D_{avg}\alpha_1}{2}\left(\ln\frac{\alpha_2}{\rho} + \alpha_3\rho + \alpha_4\rho^2\right), \tag{7.16}$$

where $D_{avg} = (D_{out} + D_{in})/2$ in Fig. 7.5 and ρ is the "fill factor" and equal to $(D_{out} - D_{in})/(D_{out} + D_{in})$ [3]. The α coefficients are chosen as follows [3]:

$$\alpha_1 = 1.27,\ \alpha_2 = 2.07,\ \alpha_3 = 0.18,\ \alpha_4 = 0.13 \text{ for square shape} \tag{7.17}$$

$$\alpha_1 = 1.07,\ \alpha_2 = 2.29,\ \alpha_3 = 0,\ \alpha_4 = 0.19 \text{ for octagonal shape.} \tag{7.18}$$

Another empirical expression is given by [3]

$$L = 1.62 \times 10^{-3} D_{out}^{-1.21} W^{-0.147} D_{avg}^{2.4} N^{1.78} S^{-0.03} \text{ for square shape} \tag{7.19}$$

$$L = 1.33 \times 10^{-3} D_{out}^{-1.21} W^{-0.163} D_{avg}^{2.43} N^{1.75} S^{-0.049} \text{ for octagonal shape.} \tag{7.20}$$

Accuracy Considerations The above inductance equations yield different levels of accuracy for different geometries. For example, the measurements on tens of inductors in [3] reveal that Eqs. (7.19) and (7.20) incur an error of about 8% for 20% of the inductors and an error of about 4% for 50% of the inductors. We must then ask: how much error is tolerable in inductance calculations? As observed throughout this book and exemplified by Fig. 7.1(b), inductors must typically resonate with their surrounding capacitances at the desired frequency. Since a small error of $\Delta L/L$ shifts the resonance frequency, ω_0, by approximately $\Delta L/(2L)$ (why?), we must determine the tolerable error in ω_0.

The resonance frequency error becomes critical in amplifiers and oscillators, but much more so in the latter. This is because, as seen abundantly in Chapter 8, the design of LC oscillators faces tight trade-offs between the "tuning range" and other parameters. Since the tuning range must encompass the error in ω_0, a large error dictates a wider tuning range, thereby degrading other aspects of the oscillator's performance. In practice, the tuning range of high-performance LC oscillators rarely exceeds $\pm 10\%$, requiring that both capacitance and inductance errors be only a small fraction of this value, e.g., a few percent. Thus, the foregoing inductance expressions may not provide sufficient accuracy for oscillator design.

Another issue with respect to inductance equations stems from the geometry limitations that they impose. Among the topologies shown in Fig. 7.9, only a few lend themselves to the above formulations. For example, the subtle differences between the structures in Figs. 7.9(b) and (c) or the parallel combination of the spirals in Fig. 7.9(f) may yield several percent of error in inductance predictions.

Another difficulty is that the inductance value also depends on the frequency of operation—albeit weakly—while most equations reported in the literature predict the low-frequency value. We elaborate on this dependence in Section 7.2.6.

Field Simulations With the foregoing sources of error in mind, how do we compute the inductance in practice? We may begin with the above approximate equations for standard structures, but must eventually resort to electromagnetic field simulations for standard or nonstandard geometries. A field simulator employs finite-element analysis to solve the steady-state field equations and compute the electrical properties of the structure at a given frequency.

A public-domain field simulator developed for analysis of inductors and transformers is called "Analysis and Simulation of Spiral Inductors and Transformers" (ASITIC) [4]. The tool can analyze a given structure and report its equivalent circuit components. While simple and efficient, ASITIC also appears to exhibit inaccuracies similar to those of the above equations [3, 5].[5]

Following rough estimates provided by formulas and/or ASITIC, we must analyze the structure in a more versatile field simulator. Examples include Agilent's "ADS," Sonnet Software's "Sonnet," and Ansoft's "HFSS." Interestingly, these tools yield slightly different values, partly due to the types of approximations that they make. For example, some do not accurately account for the thickness of the metal layers. Owing to these discrepancies, RF circuits sometimes do not exactly hit the targeted frequencies after the first fabrication, requiring slight adjustments and "silicon iterations." As a remedy, we can limit our usage to a library of inductors that have been measured and modeled carefully but at the cost of flexibility in design and layout.

7.2.4 Parasitic Capacitances

As a planar structure built upon a substrate, spiral inductors suffer from parasitic capacitances. We identify two types. (1) The metal line forming the inductor exhibits parallel-plate and fringe capacitances to the substrate [Fig. 7.12(a)]. If a wider line is chosen to reduce its resistance, then the parallel-plate component increases. (2) The adjacent turns also bear a fringe capacitance, which equivalently appears *in parallel* with each segment [Fig. 7.12(b)].

Let us first examine the effect of the capacitance to the substrate. Since in most circuits, one terminal of the inductor is at ac ground, we construct the uniformly-distributed equivalent circuit shown in Fig. 7.13, where each segment has an inductance of L_u. Our objective is to obtain a *lumped* model for this network. To simplify the analysis, we make two assumptions: (1) each two inductor segments have a mutual coupling of M, and (2)

5. In fact, Eqs. (7.19) and (7.20) have been developed based on ASITIC simulations.

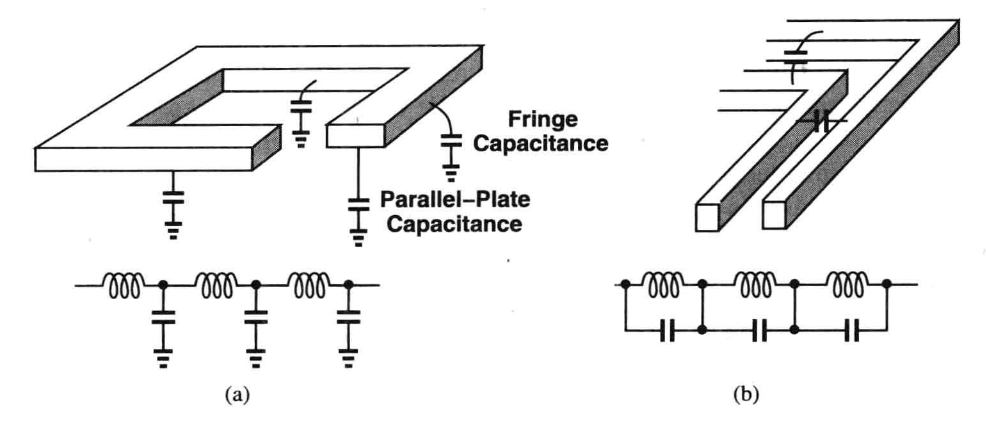


Figure 7.12 *(a) Bottom-plate and (b) interwinding capacitances of an inductor and their models.*

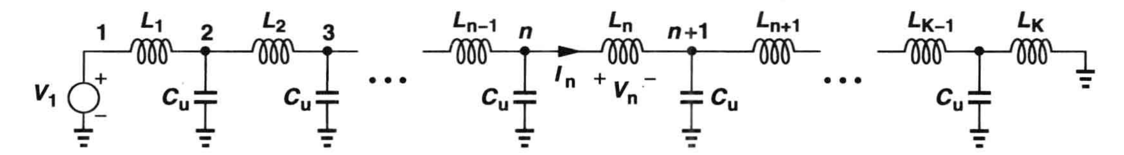


Figure 7.13 *Model of an inductor's distributed capacitance to ground.*

the coupling is strong enough that M can be assumed approximately equal to L_u. While not quite valid, these assumptions lead to a relatively accurate result.

The voltage across each inductor segment arises from the current flowing through that segment and the currents flowing through the other segments. That is,

$$V_n = j\omega L_n I_n + \sum_{m=1}^{n-1} j\omega I_m M + \sum_{m=n+1}^{K} j\omega I_m M. \tag{7.21}$$

If $M \approx L_u$, then

$$V_n = j\omega \sum_{m=1}^{K} I_m L_m. \tag{7.22}$$

Since this summation is independent of n, we note that all inductor segments sustain equal voltages [6]. The voltage at node n is therefore given by $(n/K)V_1$, yielding an electric energy stored in the corresponding node capacitance equal to

$$E_u = \frac{1}{2} C_u \left(\frac{n}{K}\right)^2 V_1^2. \tag{7.23}$$

Summing the energies stored on all of the unit capacitances, we have

$$E_{tot} = \frac{1}{2} C_u \sum_{n=1}^{K} \left(\frac{n}{K}\right)^2 V_1^2 \tag{7.24}$$

$$= \frac{1}{2} C_u \frac{(K+1)(2K+1)}{6K} V_1^2. \tag{7.25}$$

If $K \to \infty$ and $C_u \to 0$ such that KC_u is equal to the total wire capacitance, C_{tot}, then [6]

$$E_{tot} = \frac{1}{2}\frac{C_{tot}}{3}V_1^2, \tag{7.26}$$

revealing that the equivalent lumped capacitance of the spiral is given by $C_{tot}/3$ (if one end is grounded).

Let us now study the turn-to-turn (interwinding) capacitance. Using the model shown in Fig. 7.14, where $C_1 = C_2 = \cdots = C_K = C_F$, we recognize that Eq. (7.22) still applies for it is independent of capacitances. Thus, each capacitor sustains a voltage equal to V_1/K, storing an electric energy of

$$E_u = \frac{1}{2}C_F\left(\frac{1}{K}V_1\right)^2. \tag{7.27}$$

The total stored energy is given by

$$E_{tot} = KE_u \tag{7.28}$$

$$= \frac{1}{2K}C_F V_1^2. \tag{7.29}$$

Interestingly, E_{tot} falls to zero as $K \to \infty$ and $C_F \to 0$. This is because, for a large number of turns, the potential difference between adjacent turns becomes very small, yielding a small electric energy stored on the C_F's.

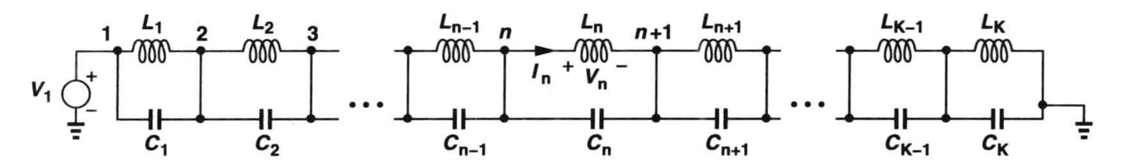


Figure 7.14 *Model of an inductor's turn-to-turn capacitances.*

In practice, we can utilize Eq. (7.29) to estimate the equivalent lumped capacitance for a finite number of turns. The following example illustrates this point.

Example 7.6

Estimate the equivalent turn-to-turn capacitance of the three-turn spiral shown in Fig. 7.15(a).

Solution:

An accurate calculation would "unwind" the structure, modeling each leg of each turn by an inductance and placing the capacitances between adjacent legs [Fig. 7.15(b)]. Unfortunately, owing to the unequal lengths of the legs, this model entails unequal inductances and capacitances, making the analysis difficult. To arrive at a *uniformly-distributed* model, we select the value of C_j equal to the average of $C_1, \ldots, C_8$, and L_j equal to the total inductance

(Continues)

Example 7.6 (*Continued*)

divided by 12. Thus, Eq. (7.29) applies and

$$C_{eq} = \frac{1}{K} C_F \tag{7.30}$$

$$= \frac{1}{8} \frac{C_1 + \cdots + C_8}{8} \tag{7.31}$$

$$= \frac{C_1 + \cdots + C_8}{64}. \tag{7.32}$$

In general, for an N-turn spiral,

$$C_{eq} = \frac{C_1 + \cdots + C_{N^2-1}}{(N^2-1)^2}. \tag{7.33}$$

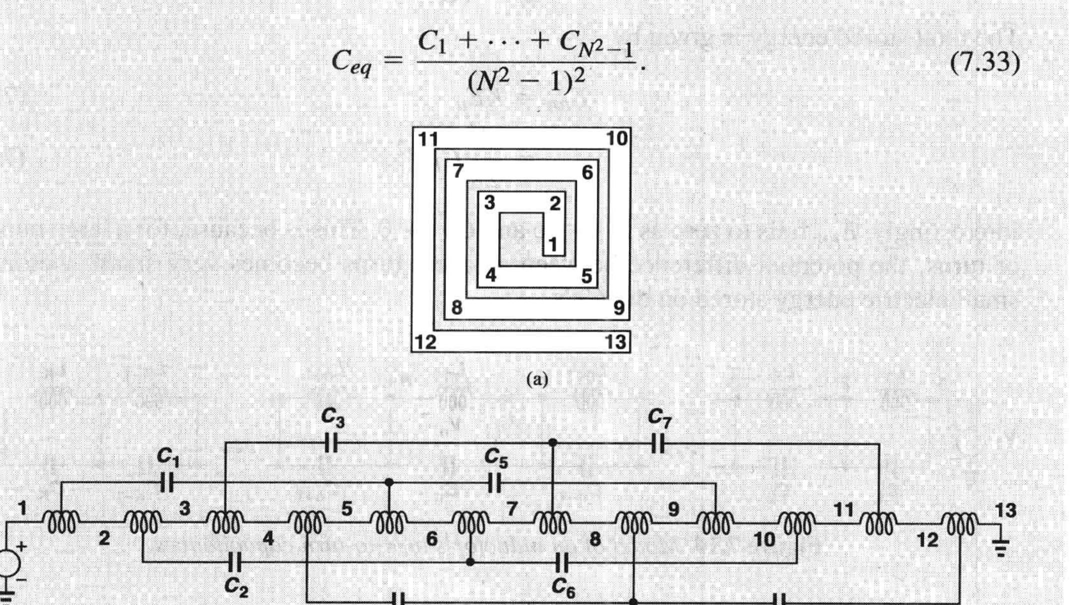


Figure 7.15 *(a) Spiral inductor for calculation of turn-to-turn capacitances, (b) circuit model.*

The frequency at which an inductor resonates with its own capacitances is called the "self-resonance frequency" (f_{SR}). In essence, the inductor behaves as a capacitor at frequencies above f_{SR}. For this reason, f_{SR} serves as a measure of the maximum frequency at which a given inductor can be used.

Example 7.7

In analogy with $Q = L\omega/R_S$ for an inductor L having a series resistance R_S, the Q of an impedance Z_1 is sometimes defined as

$$Q = \frac{Im\{Z_1\}}{Re\{Z_1\}}. \tag{7.34}$$

Example 7.7 (Continued)

Compute this Q for the parallel inductor model shown in Fig. 7.16(a).

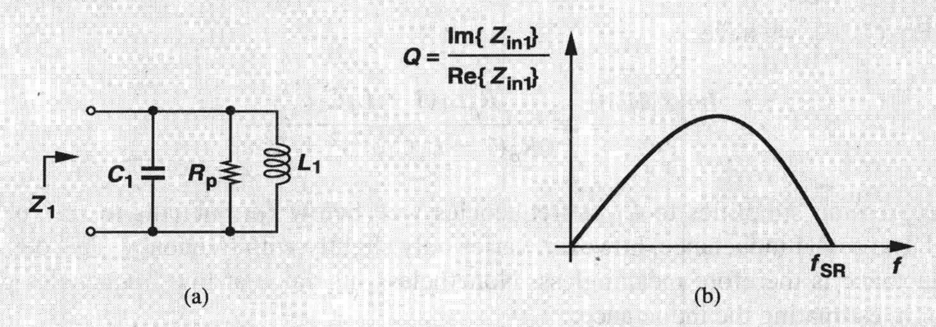


Figure 7.16 *(a) Simple tank and (b) behavior of one definition of Q.*

Solution:

We have

$$Z_1(s) = \frac{R_p L_1 s}{R_p L_1 C_1 s^2 + L_1 s + R_p}. \tag{7.35}$$

At $s = j\omega$,

$$Z_1(j\omega) = \frac{[R_p(1 - L_1 C_1 \omega^2) - jL_1\omega]jR_p L_1 \omega}{R_p^2(1 - L_1 C_1 \omega^2)^2 + L_1^2 \omega^2}. \tag{7.36}$$

It follows that

$$Q = \frac{R_p(1 - L_1 C_1 \omega^2)}{L_1 \omega} \tag{7.37}$$

$$= \frac{R_p}{L_1 \omega}\left(1 - \frac{\omega^2}{\omega_{SR}^2}\right), \tag{7.38}$$

where $\omega_{SR} = 2\pi f_{SR} = 1/\sqrt{L_1 C_1}$. At frequencies well below ω_{SR}, we have $Q \approx R_p/(L_1\omega)$, which agrees with our definition in Chapter 2. On the other hand, as the frequency approaches f_{SR}, Q falls to zero [Fig. 7.16(b)]—as if the tank were useless! This definition implies that a general impedance (including additional capacitances due to transistors, etc.) exhibits a Q of zero at resonance. Of course, the tank of Fig. 7.16(a) simply reduces to resistor R_p at f_{SR}, providing a Q of $R_p/(L_1\omega_{SR})$ rather than zero. Owing to its meaningless behavior around resonance, the Q definition given by Eq. (7.34) proves irrelevant to circuit design. We return to this point in Section 7.2.6.

Example 7.8

In analogy with L_1 = impedance/ω = $(L_1\omega)/\omega$, the equivalent inductance of a structure is sometimes defined as $Im\{Z_1(j\omega)\}/\omega$. Study this inductance definition for the parallel tank of Fig. 7.16(a) as a function of frequency.

(Continues)

Example 7.8 (*Continued*)

Solution:

From Eq. (7.36), we have

$$\frac{Im\{Z_1(j\omega)\}}{\omega} = \frac{R_p^2 L_1(1 - L_1 C_1 \omega^2)}{R_p^2(1 - L_1 C_1 \omega^2)^2 + L_1^2 \omega^2}. \tag{7.39}$$

This expression simplifies to L_1 at frequencies well below f_{SR} but falls to *zero* at resonance! The actual inductance, however, varies only slightly with frequency. This definition of inductance is therefore meaningless. Nonetheless, its value at low frequencies proves helpful in estimating the inductance.

7.2.5 Loss Mechanisms

The quality factor, Q, of inductors plays a critical role in various RF circuits. For example, the phase noise of oscillators is proportional to $1/Q^2$ (Chapter 8), and the voltage gain of "tuned amplifiers" [e.g., the CS stage in Fig. 7.1(b)] is proportional to Q. In typical CMOS technologies and for frequencies up to 5 GHz, a Q of 5 is considered moderate and a Q of 10, relatively high.

We define the Q carefully in Section 7.2.6, but for now we consider Q as a measure of how much energy is *lost* in an inductor when it carries a sinusoidal current. Since only *resistive* components dissipate energy, the loss mechanisms of inductors relate to various resistances within or around the structure that carry current when the inductor does.

In this section, we study these loss mechanisms. As we will see, it is difficult to formulate the losses analytically; we must therefore resort to simulations and even measurements to construct accurate inductor models. Nonetheless, our understanding of the loss mechanisms helps us develop guidelines for inductor modeling and design.

Metal Resistance Suppose the metal line forming an inductor exhibits a series resistance, R_S (Fig. 7.17). The Q may be defined as the ratio of the desirable impedance, $L_1\omega_0$, and the undesirable impedance, R_S:

$$Q = \frac{L_1\omega_0}{R_S}. \tag{7.40}$$

For example, a 5-nH inductor operating at 5 GHz with an R_S of 15.7 Ω has a Q of 10.

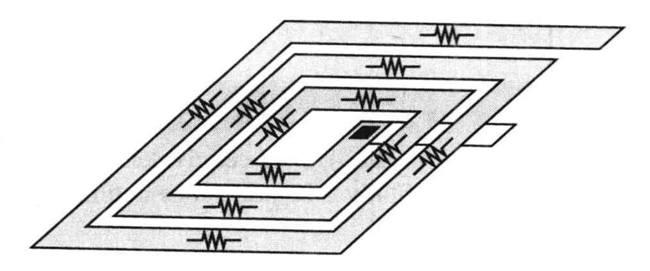


Figure 7.17 *Metal resistance in a spiral inductor.*

Example 7.9

Assuming a sheet resistance of $22\,\text{m}\Omega/\square$ for the metal, $W = 4\,\mu\text{m}$, and $S = 0.5\,\mu\text{m}$, determine if the above set of values is feasible.

Solution:

Recall from our estimates in Section 7.2.3, a 2000-μm long, 4-μm wide wire that is wound into $N = 5$ turns with $S = 0.5\,\mu\text{m}$ provides an inductance of about 4.96 nH. Such a wire consists of $2000/4 = 500$ squares and hence has a resistance of $500 \times 22\,\text{m}\Omega/\square = 11\,\Omega$. It thus appears that a Q of 10 at 5 GHz is feasible.

Unfortunately, the above example portrays an optimistic picture: the Q is limited not only by the (low-frequency) series resistance but also by several other mechanisms. That is, the overall Q may fall quite short of 10. As a rule of thumb, we strive to design inductors such that the low-frequency metal resistance yields a Q about *twice* the desired value, anticipating that other mechanisms drop the Q by a factor of 2.

How do we reduce the metal dc resistance for a given inductance? As explained in Section 7.2.3, the total length of the metal wire and the inductance are inextricably related, i.e., for a given W, S, and wire length, the inductance is a weak function of N. Thus, with W and S known, a desired inductance value translates to a certain length and hence a certain dc resistance almost regardless of the choice of N. Figure 7.18 plots the wire resistance of a 5-nH inductor with $N = 2$ to 6, $W = 4\,\mu\text{m}$, and $S = 0.5\,\mu\text{m}$. In a manner similar to the flattening effect in Fig. 7.11, R_S falls to a relatively constant value for $N > 3$.

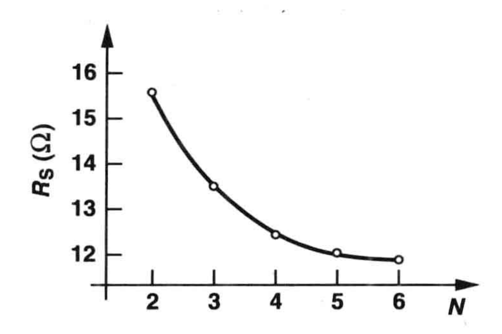


Figure 7.18 *Metal resistance of an inductor as a function of number of turns.*

From the above discussions, we conclude that the only parameter among D_{out}, S, N, and W that significantly affects the resistance is W. Of course, a wider metal line exhibits less resistance but a larger capacitance to the substrate. Spiral inductors therefore suffer from a trade-off between their Q and their parasitic capacitance. The circuit design limitations imposed by this capacitance are examined in Chapters 5 and 8.

As explained in Example 7.3, a wider metal line yields a *smaller* inductance value if S, D_{out}, and N remain constant. In other words, to retain the same inductance while W increases, we must inevitably increase D_{out} (or N), thereby increasing the length and counteracting the resistance reduction afforded by a wider line. To illustrate this effect, we can design spirals having a given inductance but different line widths and examine the

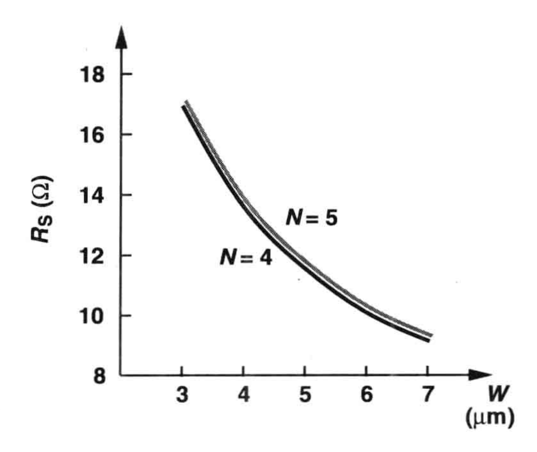


Figure 7.19 *Metal resistance of an inductor as a function of line width for different number of turns.*

resistance. Figure 7.19 plots R_S as a function of W for an inductance of 2 nH and with four or five turns. We observe that R_S falls considerably as W goes from 3 μm to about 5 μm but begins to flatten thereafter. In other words, choosing $W > 5$ μm in this example negligibly reduces the resistance but increases the parasitic capacitance proportionally.

In summary, for a given inductance value, the choice of N has little effect on R_S, and a larger W reduces R_S to some extent but at the cost of higher capacitance. These limitations manifest themselves particularly at *lower* frequencies, as shown by the following example.

Example 7.10

We wish to design a spiral inductor for a 900-MHz GSM system. Is the 5-nH structure considered in Example 7.9 suited to this application? What other choices do we have?

Solution:

Since $Q = L_1\omega_0/R_S$, if the frequency falls from 5 GHz to 900 MHz, the Q declines from 10 to 1.8.[6] Thus, a value of 5 nH is inadequate for usage at 900 MHz.

Let us attempt to raise the inductance, hoping that, in $Q = L_1\omega_0/R_S$, L_1 can increase at a higher rate than can R_S. Indeed, we observe from Eq. (7.15) that $L_1 \propto l_{tot}^{5/3}$, whereas $R_S \propto l_{tot}$. For example, if $l_{tot} = 8$ mm, $N = 10$, $W = 6$ μm, and $S = 0.5$ μm, then Eq. (7.15) yields $L \approx 35$ nH. For a sheet resistance of 22 mΩ/□, $R_S = (8000\ \mu\text{m}/6\ \mu\text{m}) \times 22\ \text{m}\Omega/\square = 29.3\ \Omega$. Thus, the Q (due to the dc resistance) reaches 6.75 at 900 MHz. Note, however, that this structure occupies a large area. The reader can readily show that the outer dimension of this spiral is approximately equal to 265 μm.

Another approach to reducing the wire resistance is to place two or more metal layers in parallel, as suggested by Fig. 7.9(f). For example, adding a metal-7 and a metal-8 spiral to a metal-9 structure lowers the resistance by about a factor of 2 because metals 7 and 8 are

6. Note that the actual Q may be even lower due to other losses.

typically half as thick as metal 9. However, the closer proximity of metal 7 to the substrate slightly raises the parasitic capacitance.

Example 7.11

A student reasons that placing m spiral inductors in parallel may in fact *degrade* the Q because it leads to an m-fold decrease in the inductance but not an m-fold decrease in resistance. Explain the flaw in the student's argument.

Solution:

Since the vertical spacing between the spirals is much less than their lateral dimensions, each two experience a strong mutual coupling (Fig. 7.20). If $L_1 = L_2 = L_3 = L$ and $M \approx L$, then the overall inductance remains equal to L (why?).

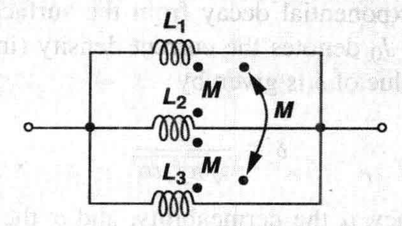


Figure 7.20 *Effect of placing tightly-coupled inductors in parallel.*

Which approach provides a more favorable resistance-capacitance trade-off: widening the metal line of a single layer or placing multiple layers in parallel? We surmise the latter; after all, if W is doubled, the capacitance of a single spiral increases by at least a factor of 2, but if metal-7 and metal-8 structures are placed in parallel with a metal-9 spiral, the capacitance may rise by only 50%. For example, the metal-9-substrate and metal-7-substrate capacitances are around 4 af/μm^2 and 6 af/μm^2, respectively. The following example demonstrates this point.

Example 7.12

Design the inductor of Example 7.10 with $W = 3\ \mu$m, $S = 0.5\ \mu$m, and $N = 10$, using metals 7, 8, and 9 in parallel.

Solution:

Since W is reduced from 6 μm to 3 μm, the term $(W + S)^{0.25}$ in the denominator of Eq. (7.15) falls by a factor of 1.17, requiring a similar drop in $l_{tot}^{5/3}$ in the numerator so as to obtain $L \approx 35$ nH. Iteration yields $l_{tot} \approx 6800\ \mu$m. The length and the outer dimension are smaller because the narrower metal line allows a tighter compaction of the turns. With three metal layers in parallel, we assume a sheet resistance of approximately 11 m$\Omega/\Box$, obtaining $R_S = 25\ \Omega$ and hence a Q of 7.9 (due to the dc resistance). The parallel combination therefore yields a higher Q.

(Continues)

Example 7.12 *(Continued)*

It is instructive to compare the capacitances of the metal-9 spiral in Example 7.10 and the above multi-layer structure. For the former, the total metal area is $l_{tot} \cdot W = 48,000\ \mu\text{m}^2$, yielding a capacitance of $(4\ \text{af}/\mu\text{m}^2) \times 48,000\ \mu\text{m}^2 = 192\,\text{fF}$.[7] For the latter, the area is equal to 20,400 μm^2 and the capacitance is 122.4 fF.

Skin Effect At high frequencies, the current through a conductor prefers to flow at the surface. If the overall current is viewed as many parallel current components, these components tend to repel each other, migrating away so as to create maximum distance between them. This trend is illustrated in Fig. 7.21. Flowing through a smaller cross section area, the high-frequency current thus faces a greater resistance. The actual distribution of the current follows an exponential decay from the surface of the conductor inward, $J(s) = J_0 \exp(-x/\delta)$, where J_0 denotes the current density (in A/m^2) at the surface, and δ is the "skin depth." The value of δ is given by

$$\delta = \frac{1}{\sqrt{\pi f \mu \sigma}}, \tag{7.41}$$

where f denotes the frequency, μ the permeability, and σ the conductivity. For example, $\delta \approx 1.4\,\mu\text{m}$ at 10 GHz for aluminum. The extra resistance of a conductor due to the skin effect is equal to

$$R_{skin} = \frac{1}{\sigma \delta}. \tag{7.42}$$

Parallel spirals can reduce this resistance if the skin depth exceeds the *sum* of the metal wire thicknesses.

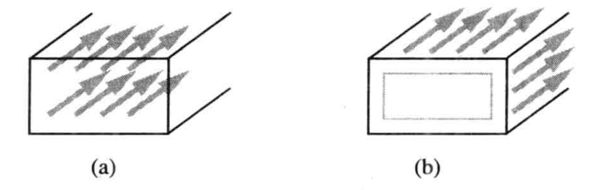


Figure 7.21 *Current distribution in a conductor at (a) low and (b) high frequencies.*

In spiral inductors, the proximity of adjacent turns results in a complex current distribution. As illustrated in Fig. 7.22(a), the current may concentrate near the edge of the wire. To understand this "current crowding" effect, consider the more detailed diagram shown in Fig. 7.22(b), where each turn carries a current of $I(t)$ [7, 8]. The current in one turn creates a time-varying magnetic field, B, that penetrates the other turns, generating loops of current.[8] Called "eddy currents," these components *add* to $I(t)$ at one edge of the wire and

7. The equivalent (lumped) capacitance of the inductor is less than this value (Section 7.2.4).

8. Faraday's law states that the voltage induced in a conducting circuit is proportional to the time derivative of the magnetic field.

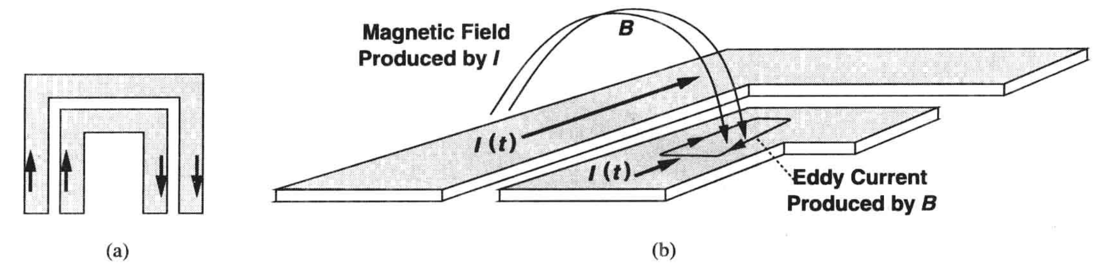


Figure 7.22 *(a) Current distribution in adjacent turns, (b) detailed view of (a).*

subtract from $I(t)$ at the other edge. Since the induced voltage increases with frequency, the eddy currents and hence the nonuniform distribution become more prominent at higher frequencies.

Based on these observations, [7, 8] derive the following expression for the resistance of a spiral inductor:

$$R_{\text{eff}} \approx R_0 \left[1 + \frac{1}{10} \left(\frac{f}{f_{crit}} \right)^2 \right], \tag{7.43}$$

where R_0 is the dc resistance and the frequency f_{crit} denotes the onset of current crowding and is given by

$$f_{crit} \approx \frac{3.1}{2\pi\mu} \frac{W+S}{W^2} R_\square. \tag{7.44}$$

In this equation, $R_\square$ represents the dc sheet resistance of the metal.

Example 7.13

Calculate the series resistance of the 30-nH inductors studied in Examples 7.9 and 7.12 at 900 MHz. Assume $\mu = 4\pi \times 10^{-7}$ H/m.

Solution:

For the single-layer spiral, $R_\square = 22\ \text{m}\Omega/\square$, $W = 6\ \mu\text{m}$, $S = 0.5\ \mu\text{m}$, and hence $f_{crit} = 1.56\ \text{GHz}$. Thus, $R_{\text{eff}} = 1.03R_0 = 30.3\ \Omega$. For the multilayer spiral, $R_\square = 11\ \text{m}\Omega/\square$, $W = 3\ \mu\text{m}$, $S = 0.5\ \mu\text{m}$, and hence $f_{crit} = 1.68\ \text{GHz}$. We therefore have $R_{\text{eff}} = 1.03R_0 = 26\ \Omega$.

Current crowding also alters the inductance and capacitance of spiral geometries. Since the current is pushed to the edge of the wire, the equivalent diameter of each turn changes slightly, yielding an inductance different from the low-frequency value. Similarly, as illustrated in Fig. 7.23(a), if a conductor carries currents only near the edges, then its middle section can be "carved out" without altering the currents and voltages, suggesting that the capacitance of this section, C_m, is immaterial. From another perspective, C_m manifests itself only if it carries displacement current, which is not possible if the middle section has no current. Based on this observation, [7, 8] approximate the total capacitance, C_{tot}, to vary

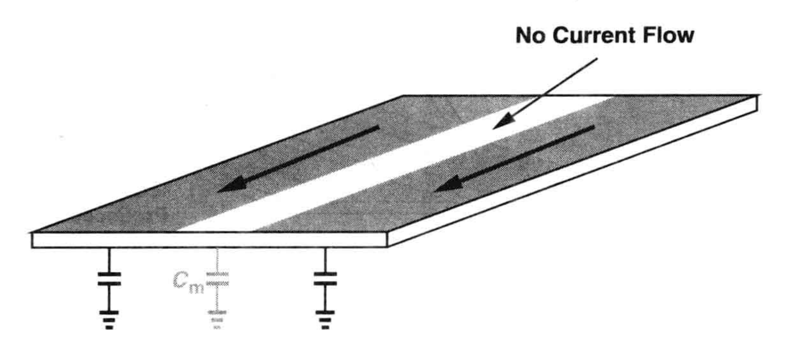


Figure 7.23 *Reduction of capacitance to the substrate as a result of current crowding.*

inversely proportional to the wire resistance:

$$C_{tot} \approx \frac{R_0}{R_{\text{eff}}} C_0, \tag{7.45}$$

where C_0 denotes the low-frequency capacitance.

Capacitive Coupling to Substrate We have seen in our studies that spirals exhibit capacitance to the substrate. As the voltage at each point on the spiral rises and falls with time, it creates a displacement current that flows through this capacitance and the substrate (Fig. 7.24). Since the substrate resistivity is neither zero nor infinity, this flow of current translates to loss in each cycle of the operation, lowering the Q.

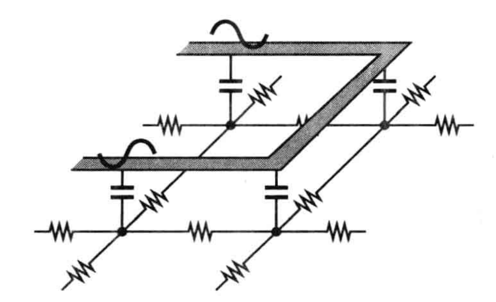


Figure 7.24 *Substrate loss due to capacitive coupling.*

Example 7.14

Use a distributed model of a spiral inductor to estimate the power lost in the substrate.

Solution:

We model the structure by K sections as shown in Fig. 7.25(a). Here, each section consists of an inductance equal to L_{tot}/K, a capacitance equal to C_{tot}/K, and a substrate resistance equal to KR_{sub}. (The other loss mechanisms are ignored here.) The factor of K in KR_{sub} is justified as follows: as we increase K for a given inductor geometry (i.e., as the distributed model approaches the actual structure), each section represents a smaller segment of the spiral and hence a smaller cross section area looking into the substrate [Fig. 7.25(b)]. Consequently, the equivalent resistance increases proportionally.

Example 7.14 (Continued)

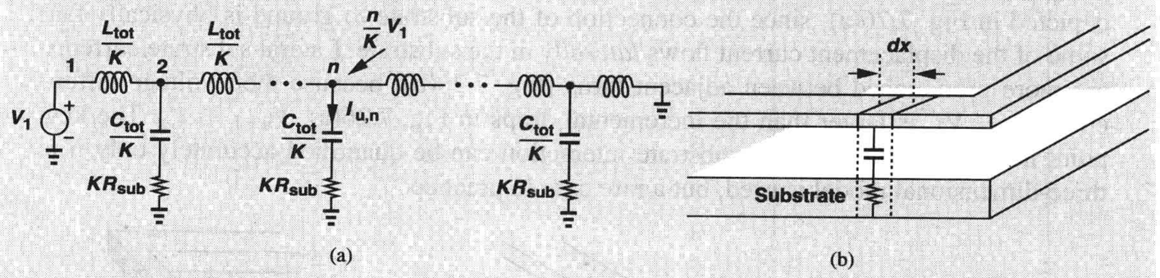


Figure 7.25 *(a) Distributed model of capacitive coupling to the substrate, (b) diagram showing an infinitesimal section.*

If we assume perfect coupling between every two inductor segments, then the voltage drop across each segment is given by Eq. (7.22):

$$V_u = \sum_{m=1}^{K} j\omega I_m L_m, \tag{7.46}$$

where I_m denotes the current flowing through segment L_m. Interestingly, due to the uniformly-distributed approximation, all segments sustain equal voltages regardless of the capacitance and resistance distribution. Thus, the voltage at node number n is given by $(n/K)V_1$ and the current flowing through the corresponding RC branch by

$$I_{u,n} = \frac{n}{K}\frac{V_1}{KR_{sub} + \left(j\frac{C_{tot}}{K}\omega\right)^{-1}}. \tag{7.47}$$

Since the average power dissipated in the resistor KR_{sub} is equal to $|I_{u,n}|^2 R_{sub}$, the *total* lost power in the spiral is obtained as

$$P_{tot} = \sum_{n=1}^{K} |I_{u,n}|^2 KR_{sub} \tag{7.48}$$

$$= \sum_{n=1}^{K} \frac{V_1^2 KR_{sub}}{K^2R_{sub}^2 + \left(\frac{C_{tot}}{K}\omega\right)^{-2}}\frac{n^2}{K^2} \tag{7.49}$$

$$= \frac{V_1^2 KR_{sub}}{K^2R_{sub}^2 + \left(\frac{C_{tot}}{K}\omega\right)^{-2}}\frac{K(K+1)(2K+1)}{6K^2}. \tag{7.50}$$

Letting K go to infinity, we have

$$P_{tot} = \frac{V_1^2}{R_{sub}^2 + (C_{tot}^2\omega^2)^{-1}}\frac{R_{sub}}{3}. \tag{7.51}$$

For example, if $R_{sub}^2 \ll (C_{tot}^2\omega^2)^{-1}$, then $P_{tot} \approx V_1^2 R_{sub} C_{tot}^2 \omega^2/3$. Conversely, if $R_{sub}^2 \gg (C_{tot}^2\omega^2)^{-1}$, then $P_{tot} \approx V_1^2/(3R_{sub})$.

The foregoing example provides insight into the power loss due to capacitive coupling to the substrate. The distributed model of the substrate, however, is not accurate. As depicted in Fig. 7.26(a), since the connection of the substrate to ground is physically far, some of the displacement current flows *laterally* in the substrate. Lateral substrate currents are more pronounced between adjacent turns [Fig. 7.26(b)] because their voltage difference, $V_1 - V_2$, is larger than the incremental drops in Fig. 7.26(a), $V_{n+1} - V_n$. The key point here is that the inductor-substrate interaction can be quantified accurately only if a three-dimensional model is used, but a rare case in practice.

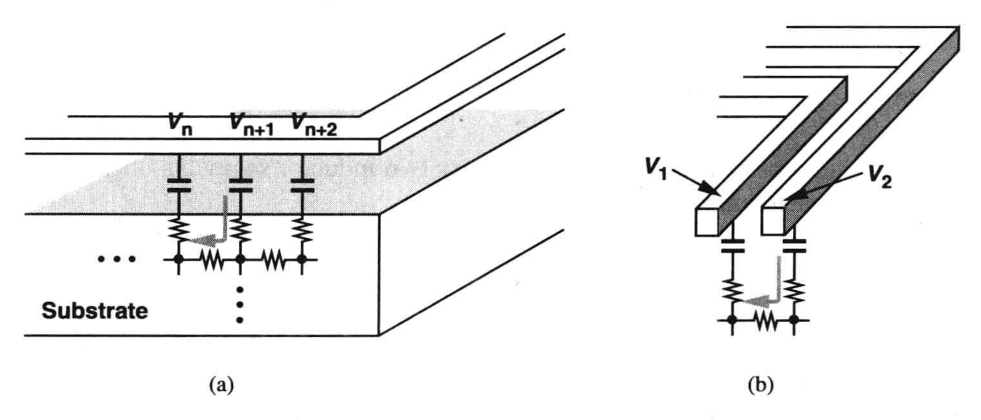


Figure 7.26 *Lateral current flow in the substrate (a) under a branch, and (b) from one branch to another.*

Magnetic Coupling to the Substrate The magnetic coupling from an inductor to the substrate can be understood with the aid of basic electromagnetic laws: (1) Ampere's law states that a current flowing through a conductor generates a magnetic field around the conductor; (2) Faraday's law states that a time-varying magnetic field induces a voltage, and hence a current if the voltage appears across a conducting material; (3) Lenz's law states that the current induced by a magnetic field generates another magnetic field opposing the first field.

Ampere's and Faraday's laws readily reveal that, as the current through an inductor varies with time, it creates an eddy current in the substrate (Fig. 7.27). Lenz's law implies that the current flows in the opposite direction. Of course, if the substrate resistance were infinity, no current would flow and no loss would occur.

The induction of eddy currents in the substrate can also be viewed as transformer coupling. As illustrated in Fig. 7.28(a), the inductor and the substrate act as the primary and

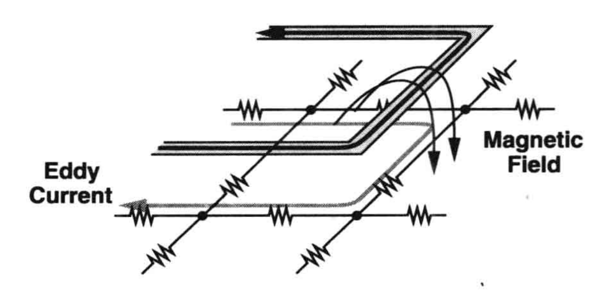


Figure 7.27 *Magnetic coupling to the substrate.*

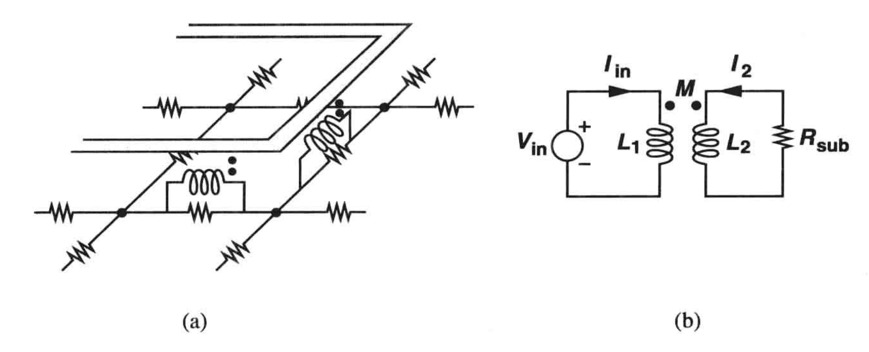


Figure 7.28 *(a) Modeling of magnetic coupling by transformers, (b) lumped model of (a).*

the secondary, respectively. Figure 7.28(b) depicts a lumped model of the overall system, with L_1 representing the spiral, M the magnetic coupling, and L_2 and R_{sub} the substrate. It follows that

$$V_{in} = L_1 s I_{in} + M s I_2 \tag{7.52}$$

$$-R_{sub} I_2 = I_2 L_2 s + M s I_{in}. \tag{7.53}$$

Thus,

$$\frac{V_{in}}{I_{in}} = L_1 s - \frac{M^2 s^2}{R_{sub} + L_2 s}. \tag{7.54}$$

For $s = j\omega$,

$$\frac{V_{in}}{I_{in}} = \frac{M^2\omega^2 R_{sub}}{R_{sub}^2 + L_2^2\omega^2} + \left(L_1 - \frac{M^2\omega^2 L_2}{R_{sub}^2 + L_2^2\omega^2}\right) j\omega, \tag{7.55}$$

implying that R_{sub} is transformed by a factor of $M^2\omega^2/(R_{sub}^2 + L_2^2\omega^2)$ and the inductance is *reduced* by an amount equal to $M^2\omega^2 L_2/(R_{sub}^2 + L_2^2\omega^2)$.

Example 7.15

A student concludes that both the electric coupling and the magnetic coupling to the substrate are eliminated if a grounded conductive plate is placed under the spiral (Fig. 7.29). Explain the pros and cons of this approach.

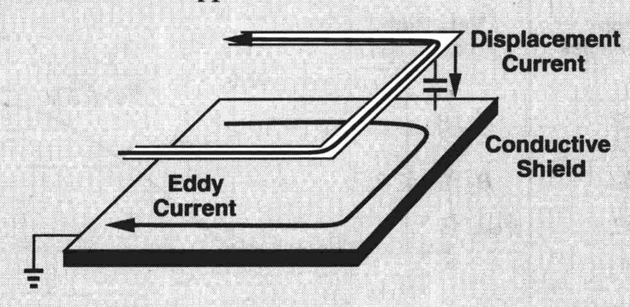


Figure 7.29 *Inductor with a continuous shield.*

(Continues)

Example 7.15 (*Continued*)

Solution:

This method indeed reduces the path resistance seen by both displacement and eddy currents. Unfortunately, however, Eq. (7.55) reveals that the equivalent inductance also falls with R_{sub}. For $R_{sub} = 0$,

$$L_{eq} = L_1 - \frac{M^2}{L_2}. \tag{7.56}$$

Since the vertical spacing between the spiral and the conductive plate ($\approx 5\ \mu$m) is much smaller than their lateral dimensions, we have $M \approx L_2 \approx L_1$, obtaining a very small value for L_{eq}. In other words, even though the substrate losses are reduced, the drastic fall in the equivalent inductance still yields a low Q because of the spiral's resistance.

It is instructive to consider a few special cases of Eq. (7.54). If $L_1 = L_2 = M$, then

$$\frac{V_{in}}{I_{in}} = L_1 s || R_{sub}, \tag{7.57}$$

indicating that R_{sub} simply appears in parallel with L_1, lowering the Q.

Example 7.16

Sketch the Q of a given inductor as a function of frequency.

Solution:

At low frequencies, the Q is given by the dc resistance of the spiral, R_S. As the frequency increases, $Q = L_1\omega/R_S$ rises linearly up to a point where skin effect becomes significant [Fig. 7.30(a)]. The Q then increases in proportion to $\sqrt{f}$. At higher frequencies, $L_1\omega \gg R_S$, and Eq. (7.57) reveals that R_{sub} shunts the inductor, limiting the Q to

$$Q \approx \frac{R_{sub}}{L_1\omega}, \tag{7.58}$$

which *falls* with frequency. Figure 7.30(b) sketches the behavior.

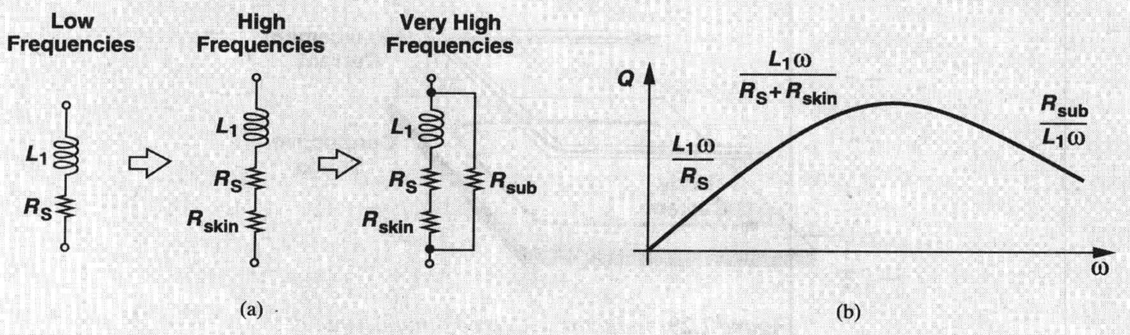


Figure 7.30 *(a) Inductor model reflecting loss at different frequencies, (b) corresponding Q behavior.*

As another special case, suppose $R_{sub} \ll |L_2 s|$. We can then factor $L_2 s$ out in Eq. (7.54) and approximate the result as

$$\frac{V_{in}}{I_{in}} = \left(L_1 - \frac{M^2}{L_2}\right)s + \frac{M^2}{L_2^2}R_{sub}. \tag{7.59}$$

Thus, as predicted in Example 7.15, the inductance is reduced by an amount equal to M^2/L_2. Moreover, the substrate resistance is transformed by a factor of M^2/L_2^2 and appears in *series* with the net inductance.

7.2.6 Inductor Modeling

Our study of various effects in spiral inductors has prepared us for developing a circuit model that can be used in simulations. Ideally, we wish to obtain a model that retains our physical insights and is both simple and accurate. In practice, some compromise must be made.

It is important to note that (1) both the spiral and the substrate act as three-dimensional distributed structures and can only be approximated by a two-dimensional lumped model; (2) due to skin effect, current-crowding effects, and eddy currents, some of the inductor parameters vary with frequency, making it difficult to fit the model in a broad bandwidth.

Example 7.17

If RF design mostly deals with narrowband systems, why is a broadband model necessary?

Solution:

From a practical point of view, it is desirable to develop a broadband model for a given inductor structure so that it can be used by multiple designers and at different frequencies without repeating the modeling effort each time. Moreover, RF systems such as ultra-wideband (UWB) and cognitive radios do operate across a wide bandwidth, requiring broadband models.

Let us begin with a model representing metal losses. As shown in Fig. 7.31(a), a series resistance can embody both low-frequency and skin resistance. With a constant R_S, the model is valid for a limited frequency range. As explained in Chapter 2, the loss can alternatively be modeled by a parallel resistance [Fig. 7.31(b)] but still for a narrow range if R_p is constant.

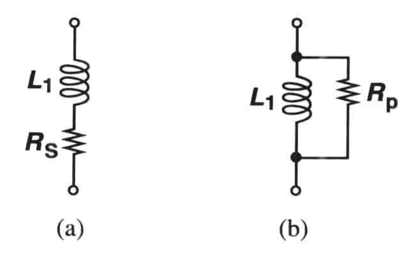


Figure 7.31 *Modeling loss by (a) series or (b) parallel resistors.*

An interesting observation allows us to combine the models of Figs. 7.31(a) and (b), thus broadening the valid bandwidth. The following example serves as the starting point.

Example 7.18

If the inductance and resistance values in Fig. 7.31 are independent of frequency, how do the two models predict the behavior of the Q?

Solution:

In Fig. 7.31(a), $Q = L_1\omega/R_S$, whereas in Fig. 7.31(b), $Q = R_p/(L_1\omega)$; i.e., the two models predict opposite trends with frequency. (We also encountered this effect in Example 7.16.)

The above observation suggests that we can tailor the frequency dependence of the Q by merging the two models. Depicted in Fig. 7.32(a), such a model partitions the loss between a series resistance and a parallel resistance. A simple approach assigns half of the loss to each at the center frequency of the band:

$$R_S' = \frac{L_1\omega}{2Q} \tag{7.60}$$

$$R_p' = 2QL_1\omega. \tag{7.61}$$

In Problem 7.2, we prove that the overall Q of the circuit, defined as $Im\{Z_1\}/Re\{Z_1\}$, is equal to

$$Q = \frac{L_1\omega R_p'}{L_1^2\omega^2 + R_S'(R_S' + R_p')}. \tag{7.62}$$

Note that this definition of Q is meaningful here because the circuit does not resonate at any frequency. As shown in Fig. 7.32(b), the Q reaches a peak of $2\sqrt{R_p'/R_S'}$ at $\omega_0 = \sqrt{R_S'R_p'}/L_1$. The choice of R_S' and R_p' can therefore yield an accurate variation for a certain frequency range.

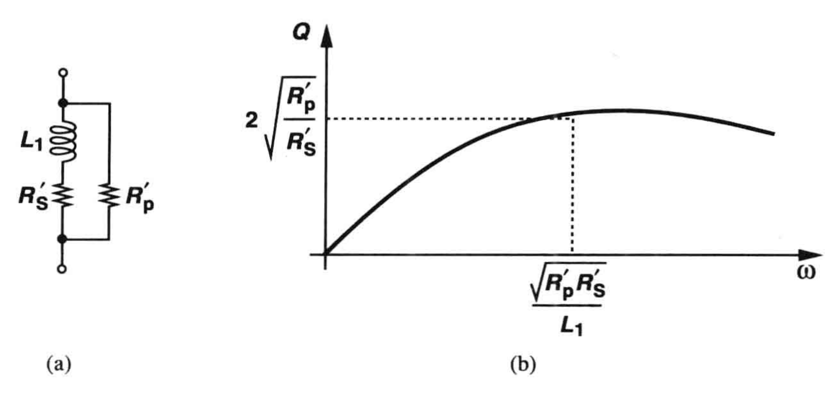


Figure 7.32 *(a) Modeling loss by both series and parallel resistors, (b) resulting Q behavior.*

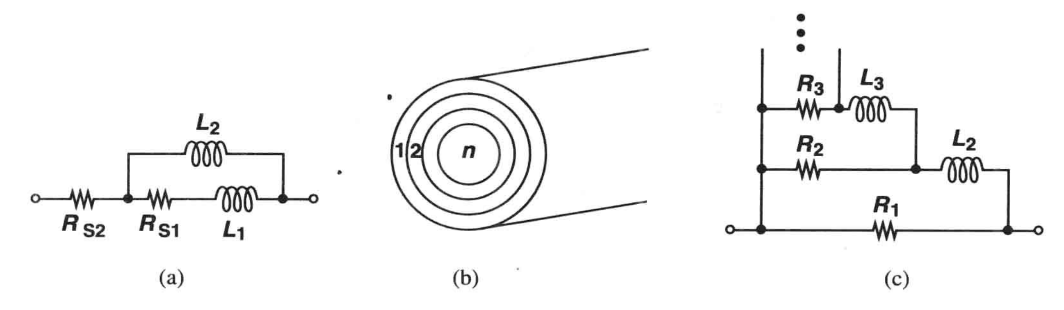


Figure 7.33 *(a) Broadband model of inductor, (b) view of a conductor as concentric cylinders, (c) broadband skin effect model.*

A more general model of skin effect has been proposed by [9] and is illustrated in Fig. 7.33. Suppose a model must be valid only at dc and a high frequency. Then, as shown in Fig. 7.33(a), we select a series resistance, R_{S1}, equal to that due to skin effect and shunt the combination of R_{S1} and L_1 with a large inductor, L_2. We then add R_{S2} in series to model the low-frequency resistance of the wire. At high frequencies, L_2 is open and $R_{S1} + R_{S2}$ embodies the overall loss; at low frequencies, the network reduces to R_{S2}.

The above principle can be extended to broadband modeling of skin effect. Depicted in Fig. 7.33(b) for a cylindrical wire, the approach in [9] views the line as a set of concentric cylinders, each having some low-frequency resistance and inductance, arriving at the circuit in Fig. 7.33(c) for one section of the distributed model. Here, the branch consisting of R_j and L_j represents the impedance of cylinder number j. At low frequencies, the current is uniformly distributed through the conductor and the model reduces to $R_1||R_2||\cdots||R_n$ [9]. As the frequency increases, the current moves away from the inner cylinders, as modeled by the rising impedance of the inductors in each branch. In [9], a constant ratio R_j/R_{j+1} is maintained to simplify the model. We return to the use of this model for inductors later in this section.

We now add the effect of capacitive coupling to the substrate. Figure 7.34(a) shows a one-dimensional uniformly-distributed model where the total inductance and series resistance are decomposed into n equal segments, i.e., $L_1 + L_2 + \cdots + L_n = L_{tot}$ and $R_{S1} + R_{S2} + \cdots + R_{Sn} = R_{S,tot}$.[9] The nodes in the substrate are connected to one another by $R_{sub1}, \ldots, R_{sub,n-1}$ and to ground by $R_{G1}, \ldots, R_{Gn}$. The total capacitance between the spiral and the substrate is decomposed into $C_{sub1}, \ldots, C_{subn}$.

Continuing our model development, we include the magnetic coupling to the substrate. As depicted in Fig. 7.34(b), each inductor segment is coupled to the substrate through a transformer. Proper choice of the mutual coupling and R_{subm} allows accurate representation of this type of loss. In this model, the capacitance between the substrate nodes is also included.

While capturing the physical properties of inductors, the model shown in Fig. 7.34(b) proves too complex for practical use. The principal issue is that the numerous parameters make it difficult to fit the model to measured data. We must therefore seek more compact models that more easily lend themselves to parameter extraction and fitting. In the first

9. A more accurate model would include mutual coupling such that $L_{tot} = L_1 + \cdots + L_n + nM$.

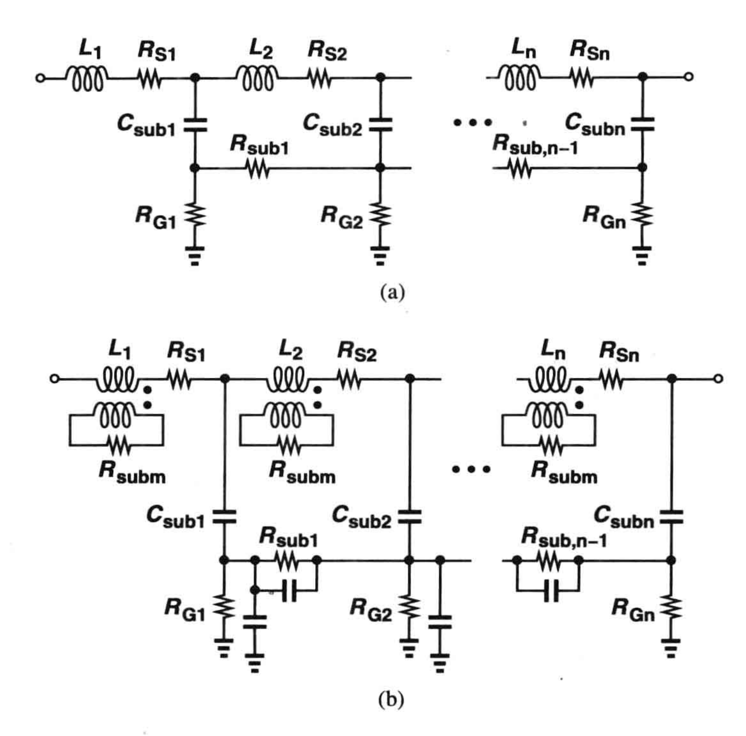


Figure 7.34 *Distributed inductor model with (a) capacitive and (b) magnetic coupling to substrate.*

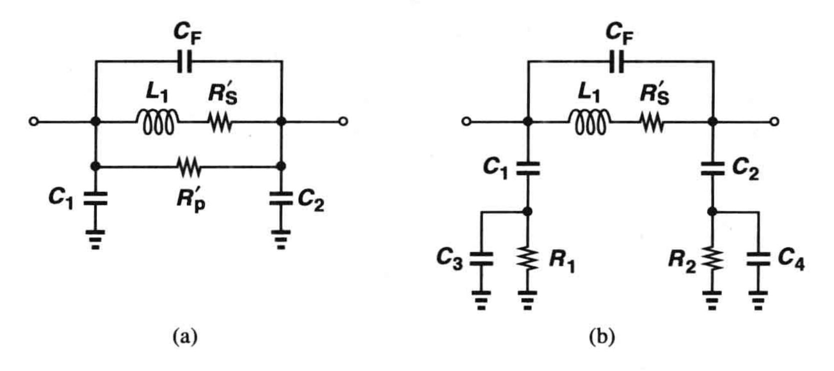


Figure 7.35 *(a) Compact inductor model, (b) alternative topology.*

step, we turn to lumped models. As a simple example, we return to the parallel-series combination of Fig. 7.32(a) and add capacitances to the substrate [Fig. 7.35(a)]. We surmise that R_S' and R_p' can represent *all* of the losses even though they do not physically reflect the substrate loss. We also recall from Section 7.2.4 that an equivalent lumped capacitance, C_F, appears between the two terminals. With constant element values, this model is accurate for a bandwidth of about ±20% around the center frequency.

An interesting dilemma arises in the above lumped model. We may choose C_1 and C_2 to be equal to half of the total capacitance to the substrate, but our analysis in Section 7.2.4 suggests that, if one terminal is grounded, the equivalent capacitance is one-*third* of the total amount. This is a shortcoming of the lumped model.

Another model that has proved relatively accurate is shown in Fig. 7.35(b). Here, R_1 and R_2 play a similar role to that of R_p in Fig. 7.35(a). Note that neither model explicitly includes the magnetic coupling to the substrate. The assumption is that the three resistances suffice to represent all of the losses across a reasonable bandwidth (e.g., ±20% around the frequency at which the component values are calculated). A more broadband model is described in [10].

Definitions of Q In this book, we have encountered several definitions of the Q of an inductor:

$$Q_1 = \frac{L\omega}{R_S} \tag{7.63}$$

$$Q_2 = \frac{R_p}{L\omega} \tag{7.64}$$

$$Q_3 = \frac{Im\{Z\}}{Re\{Z\}}. \tag{7.65}$$

In basic physics, the Q of a lossy oscillatory system is defined as

$$Q_4 = 2\pi \frac{\text{Energy Stored}}{\text{Energy Dissipated per Cycle}}. \tag{7.66}$$

Additionally, for a second-order tank, the Q can be defined in terms of the resonance frequency, ω_0, and the -3-dB bandwidth, ω_{BW}, as

$$Q_5 = \frac{\omega_0}{\omega_{BW}}. \tag{7.67}$$

To make matters more complicated, we can also define the Q of an open-loop system at a frequency ω_0 as

$$Q_6 = \frac{\omega_0}{2}\frac{d\phi}{d\omega}, \tag{7.68}$$

where ϕ denotes the phase of the system's transfer function (Chapter 8).

Which one of the above definitions is relevant to RF design? We recall from Chapter 2 that Q_1 and Q_2 model the loss by a single resistance and are equivalent for a narrow bandwidth. Also, from Example 7.7, we discard Q_3 because it fails where it matters most: in most RF circuits, inductors operate in resonance (with their own and other circuit capacitances), exhibiting $Q_3 = 0$. The remaining three, namely, Q_4, Q_5, and Q_6, are equivalent for a second-order tank in the vicinity of the resonance frequency.

Before narrowing down the definitions of Q further, we must recognize that, in general, the analysis of a circuit does *not* require a knowledge of the Q's of its constituent devices. For example, the inductor model shown in Fig. 7.34(b) represents the properties of the device completely. Thus, the concept of Q has been invented primarily to provide intuition, allowing analysis by inspection as well as the use of certain rules of thumb.

In this book, we mostly deal with only one of the above definitions, Q_2. We reduce any resonant network to a parallel RLC tank, lumping all of the loss in a single parallel resistor R_p, and define $Q_2 = R_P/(L\omega_0)$. This readily yields the voltage gain of the stage

shown in Fig. 7.1(b) as $-g_m(r_O||R_p)$ at resonance. Moreover, if we wish to compute the Q of a given inductor design at *different* frequencies, then we add or subtract enough parallel capacitance to create resonance at each frequency and determine Q_2 accordingly.

It is interesting to note the following equivalencies for a second-order parallel tank: for Q_2 and Q_3, we have

$$Q_2 = 2\pi \frac{\text{Peak Magnetic Energy}}{\text{Energy Lost per Cycle}} \tag{7.69}$$

$$Q_3 = 2\pi \frac{\text{Peak Magnetic Energy} - \text{Peak Electric Energy}}{\text{Energy Lost per Cycle}}. \tag{7.70}$$

7.2.7 Alternative Inductor Structures

As illustrated conceptually in Fig. 7.9, many variants of spiral inductors can be envisioned that can potentially raise the Q, lower the parasitic capacitances, or reduce the lateral dimensions. For example, the parallel combination of spirals proves beneficial in reducing the metal resistance. In this section, we deal with several inductor geometries.

Symmetric Inductors Differential circuits can employ a single symmetric inductor rather than two (asymmetric) spirals (Fig. 7.36). In addition to saving area, a differential geometry (driven by differential signals) also exhibits a higher Q [11]. To understand this property, let us use the model of Fig. 7.35(b) with single-ended and differential stimuli (Fig. 7.37). If in Fig. 7.37(a), we neglect C_3 and assume C_1 has a low impedance, then the resistance shunting the inductor at high frequencies is approximately equal to R_1. That is, the circuit is reduced to that in Fig. 7.37(b).

Now, consider the differential arrangement shown in Fig. 7.37(c). The circuit can be decomposed into two symmetric half circuits, revealing that R_1 (or R_2) appears in parallel with an inductance of $L/2$ [Fig. 7.37(d)] and hence affects the Q to a lesser extent [11]. In Problem 7.4, we use Eq. (7.62) to compare the Q's in the two cases. For frequencies above 5 GHz, differential spirals provide a Q of 8 or higher and single-ended structures a Q of about 5 to 6.

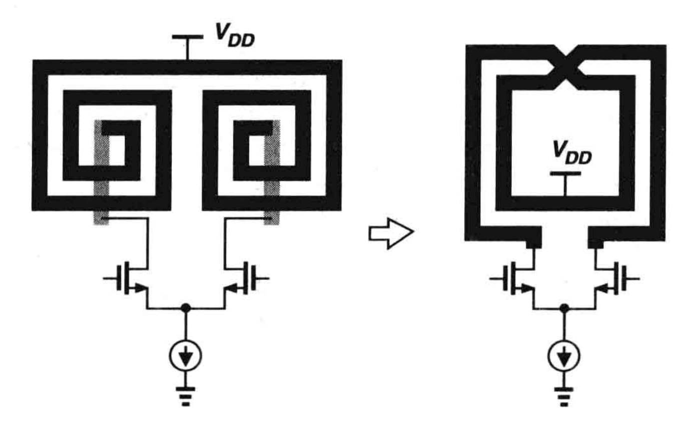


Figure 7.36 *Use of symmetric inductor in a differential circuit.*

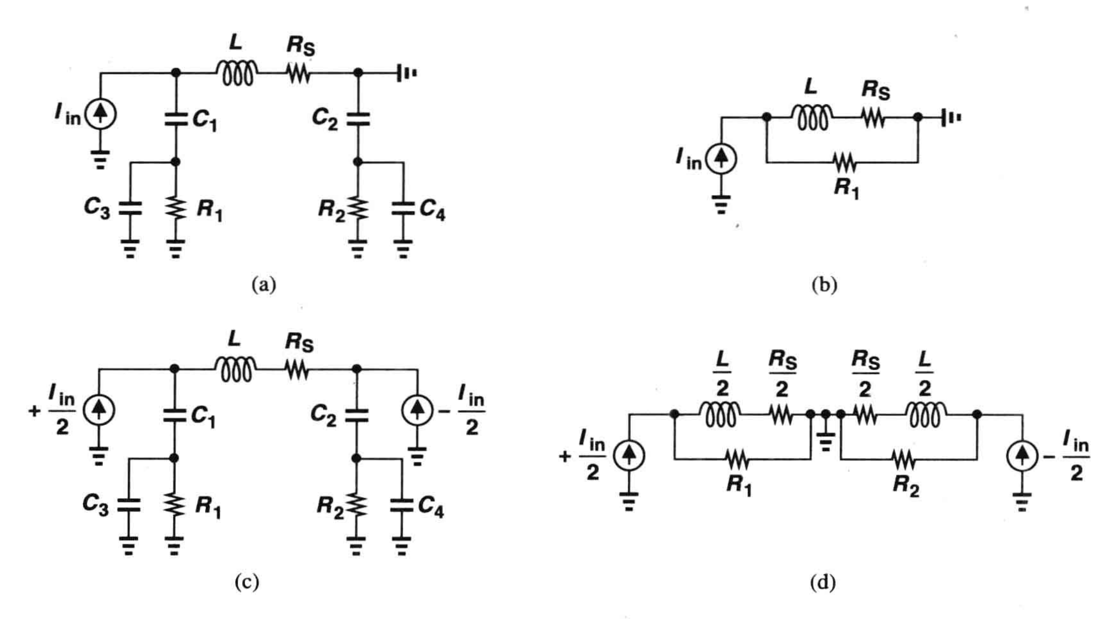


Figure 7.37 *(a) Inductor driven by a single-ended input, (b) simplified model of (a), (c) symmetric inductor driven by differential inputs, (d) simplified model of (c).*

The principal drawback of symmetric inductors is their large interwinding capacitance, a point of contrast to the trend predicted by Eq. (7.29). Consider the arrangement shown in Fig. 7.38(a), where the inductor is driven by differential voltages and viewed as four segments in series. Modeling each segment by an inductor and including the fringe capacitance between the segments, we obtain the network depicted in Fig. 7.38(b). Note that symmetry creates a virtual ground at node 3. This model implies that C_1 and C_2 sustain large voltages, e.g., as much as $V_{in}/2$ if we assume a linear voltage profile from node 1 to node 5 [Fig. 7.38(c)].

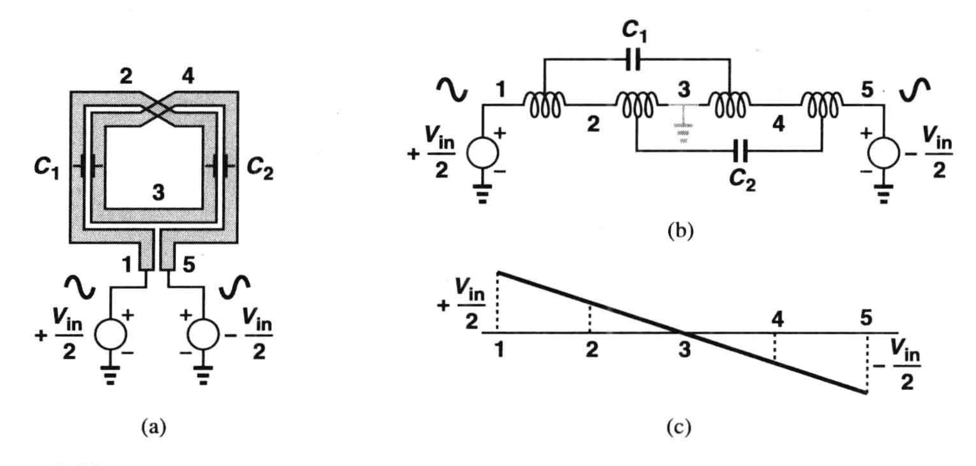


Figure 7.38 *(a) Symmetric inductor, (b) equivalent circuit, (c) voltage profile along the inductor.*

Example 7.19

Estimate the equivalent lumped interwinding capacitance of the three-turn spiral shown in Fig. 7.39(a).

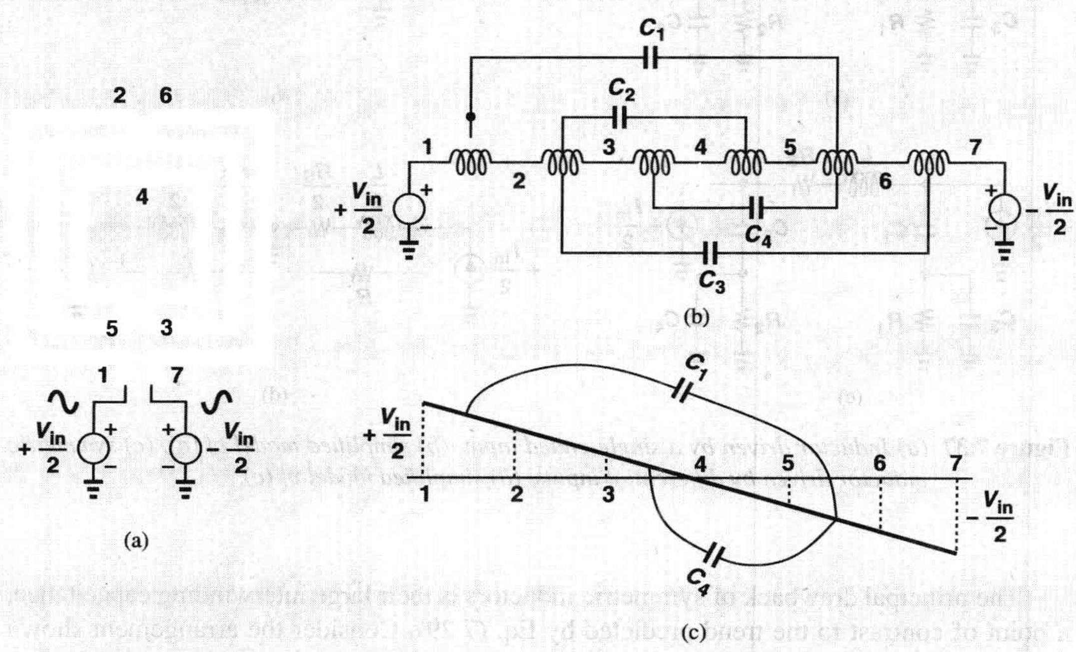


Figure 7.39 *(a) Three-turn symmetric inductor, (b) equivalent circuit, (c) voltage profile along the ladder.*

Solution:

We unwind the structure as depicted in Fig. 7.39(b), assuming, as an approximation, that all unit inductances are equal and so are all unit capacitances. We further assume a linear voltage profile from one end to the other [Fig. 7.39(c)]. Thus, C_1 sustains a voltage of $4V_{in}/6$ and so does C_3. Similarly, each of C_2 and C_4 has a voltage of $2V_{in}/6$. The total electric energy stored on the four capacitors is therefore equal to

$$E_{tot} = 2\left[\frac{1}{2}C\left(\frac{2}{3}V_{in}\right)^2 + \frac{1}{2}C\left(\frac{1}{3}V_{in}\right)^2\right], \tag{7.71}$$

where $C = C_1 = \cdots = C_4$. Denoting $C_1 + \cdots + C_4$ by C_{tot}, we have

$$E_{tot} = \frac{5}{9}\frac{C_{tot}}{4}V_{in}^2, \tag{7.72}$$

and hence an equivalent lumped capacitance of

$$C_{eq} = \frac{5}{18}C_{tot}. \tag{7.73}$$

Example 7.19 *(Continued)*

Compared with its counterpart in a single-ended inductor, Eq. (7.32), this value is higher by a factor of $160/9 \approx 18$. In fact, the equivalent interwinding capacitance of a differential inductor is typically quite *larger* than the capacitance to the substrate, dominating the self-resonance frequency.

How do we reduce the interwinding capacitance? We can increase the line-to-line spacing, S, but, for a given outer dimension, this results in smaller inner turns and hence a lower inductance. In fact, Eq. (7.15) reveals that L falls as S increases and l_{tot} remains constant, yielding a lower Q. As a rule of thumb, we choose a spacing of approximately three times the minimum allowable value.[10] Further increase of S lowers the fringe capacitance only slightly but degrades the Q.

Owing to their higher Q, differential inductors are common in oscillator design, where the Q matters most. They are typically constructed as octagons (a symmetric version of that in Fig. 7.9(b)] because, for a given inductance, an octagonal shape has a shorter length and hence less series resistance than does a square geometry. (Perpendicular sides provide little mutual coupling.) For other differential circuits, such structures can be used, but at the cost of routing complexity. Figure 7.40 illustrates this point for a cascade of two stages. With single-ended spirals on each side, the lines traveling to the next stage can pass *between* the inductors [Fig. 7.40(a)]. Of course, some spacing is necessary between the lines and the inductors so as to minimize unwanted coupling. On the other hand, with the differential structure, the lines must travel either *through* the inductor or *around* it [Fig. 7.40(b)], creating greater coupling than in the former case.

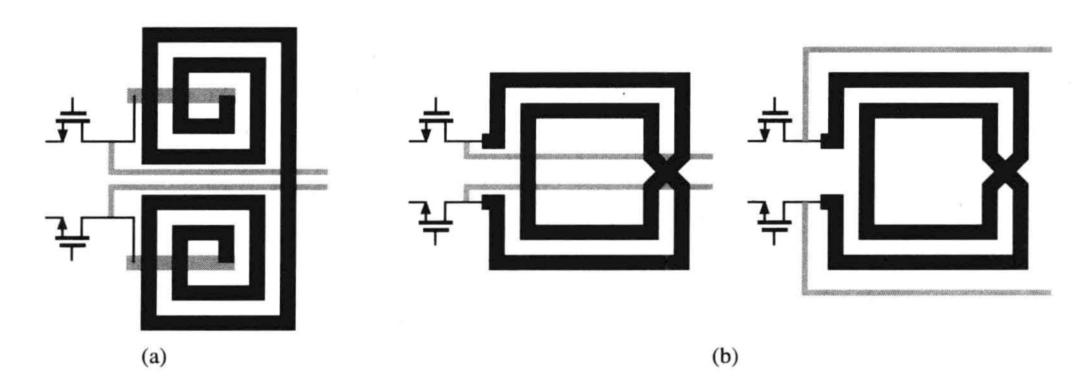


Figure 7.40 *Routing of signals to next stage in a circuit using (a) single-ended inductors, (b) a symmetric inductor.*

10. But, in some technologies long lines require a wider spacing than short lines, in which case the minimum S may be 1 to 1.5 μm.

Example 7.20

If used as the load of differential circuits, single-ended inductors can be laid out with "mirror symmetry" [Fig. 7.41(a)] or "step symmetry" [Fig. 7.41(b)]. Discuss the pros and cons of each layout style.

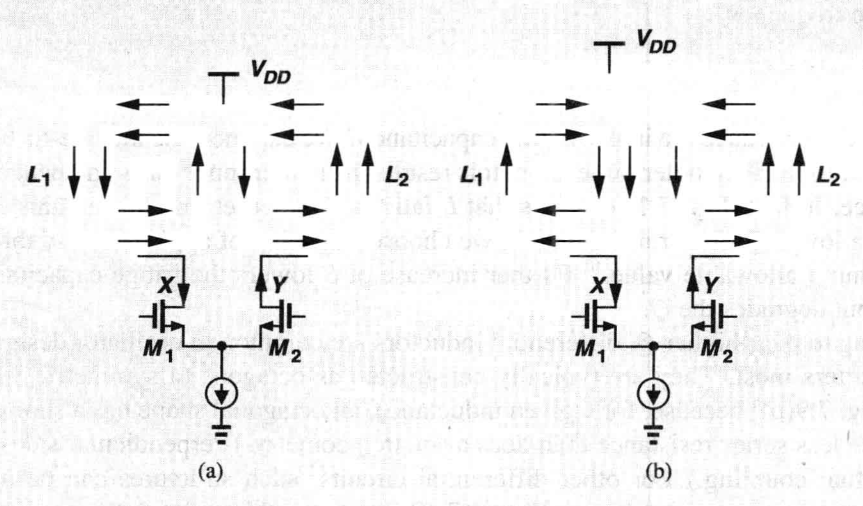


Figure 7.41 *Load inductors in a differential pair with (a) mirror symmetry and (b) step symmetry.*

Solution:

The circuit of Fig. 7.41(a) is relatively symmetric but suffers from *undesirable* mutual coupling between L_1 and L_2. Since the differential currents produced by M_1 and M_2 flow in *opposite* directions in the spirals, the equivalent inductance seen between X and Y is equal to

$$L_{eq} = L_1 + L_2 - 2M, \tag{7.74}$$

where M denotes the mutual coupling between L_1 and L_2. With a small spacing between the spirals, the mutual coupling factor, k, may reach roughly 0.25, yielding $M = k\sqrt{L_1 L_2} = 0.25L$ if $L_1 = L_2 = L$. In other words, L_{eq} is 25% less than $L_1 + L_2$, exhibiting a lower Q. For k to fall to a few percent, the spacing between L_1 and L_2 must exceed approximately one-half of the outer dimension of each.

In the topology of Fig. 7.41(b), the direction of currents results in

$$L_{eq} = L_1 + L_2 + 2M, \tag{7.75}$$

increasing the Q. However, the circuit is less symmetric. Thus, if symmetry is critical [e.g., in the LO buffer of a direct-conversion receiver (Chapter 4)], then we choose the former with some spacing between L_1 and L_2. Otherwise, we opt for the latter.

Another important difference between two single-ended inductors and one differential inductor is the amount of signal coupling that they inflict or incur. Consider the topology of Fig. 7.42(a) and a point P on its axis of symmetry. Using the right-hand rule, we observe that the magnetic field due to L_1 points into the page at P and that due to L_2 out of the page.

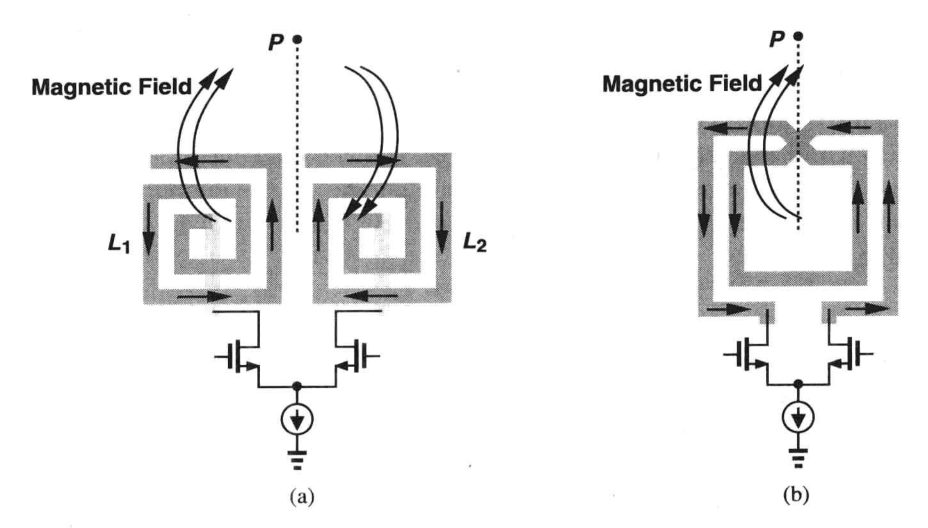


Figure 7.42 *Magnetic coupling along the axis of symmetry with (a) single-ended inductors and (b) a symmetric inductor.*

The two fields therefore cancel along the axis of symmetry. By contrast, the differential spiral in Fig. 7.42(b) produces a single magnetic field at P and hence coupling to other devices even on the line of symmetry.[11] This issue is particularly problematic in oscillators: to achieve a high Q, we wish to use symmetric inductors but at the cost of making the circuit more sensitive to injection-pulling by the power amplifier.

Example 7.21

The topology of Fig. 7.43 may be considered a candidate for small coupling. Explain the pros and cons of this structure.

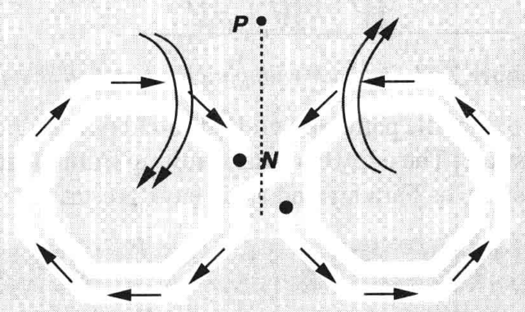


Figure 7.43 *Inductor with reduced magnetic coupling along axis of symmetry.*

Solution:

This geometry in fact consists of two single-ended inductors because node N is a virtual ground. The magnetic fields of the two halves indeed cancel on the axis of symmetry.

(*Continues*)

11. One can also view the single spiral as a loop antenna.

Example 7.21 *(Continued)*

The structure is more symmetric than the single-ended spirals with step symmetry in Fig. 7.42(a). Unfortunately, the Q of this topology is lower than that of a differential inductor because each half experiences its own substrate loss; i.e., the doubling of the substrate shunt resistance observed in Fig. 7.37 does not occur here. A variant of this structure is described in [12].

Inductors with Ground Shield In our early study of substrate loss in Section 7.2.5, we contemplated the use of a grounded shield below the inductor. The goal was to allow the displacement current to flow through a low resistance to ground, thus avoiding the loss due to electric coupling to the substrate. But we observed that eddy currents in a continuous shield drastically reduce the inductance and the Q.

We now observe that the shield can provide a low-resistance termination for electric field lines even if it is *not* continuous. As illustrated in Fig. 7.44 [13], a "patterned" shield, i.e., a plane broken periodically in the direction perpendicular to the flow of eddy currents, receives most of the electric field lines without reducing the inductance. A small fraction of the field lines sneak through the gaps in the shield and terminate on the lossy substrate. Thus, the width of the gaps must be minimized.

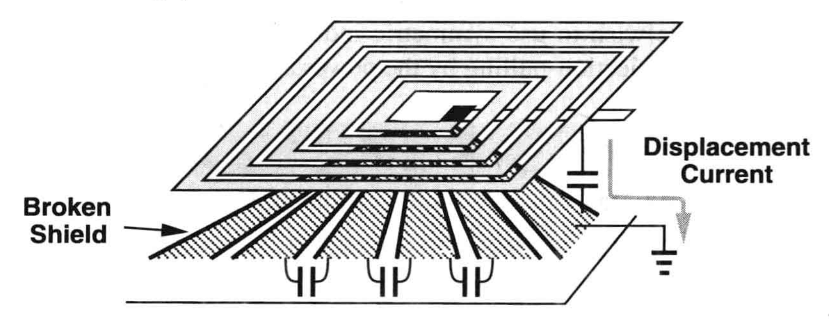


Figure 7.44 *Inductor with patterned ground shield.*

It is important to note that the patterned ground shield only reduces the effect of capacitive coupling to the substrate. The eddy currents resulting from magnetic coupling continue to flow through the substrate as Faraday and Lenz have prescribed.

Example 7.22

A student designing a patterned ground shield decides that minimizing the gap width is not a good idea because it increases the capacitance between each two sections of the shield, potentially allowing large eddy currents to flow through the shield. Is the student correct?

Solution:

While it is true that the gap capacitance increases, we must note that all of the gap capacitances appear in *series* with the path of eddy currents. The overall equivalent capacitance is therefore very small and the impedance presented to eddy currents quite high.

The use of a patterned shield may increase the Q by 10 to 15% [13], but this improvement depends on many factors and has thus been inconsistent in different reports [14]. The factors include single-ended versus differential operation, the thickness of the metal, and the resistivity of the substrate. The improvement comes at the cost of higher capacitance. For example, if the inductor is realized in metal 9 and the shield in metal 1, then the capacitance rises by about 15%. One can utilize a patterned n^+ region in the substrate as the shield to avoid this capacitance increase, but the measurement results have not been consistent.

The other difficulty with patterned shields is the additional complexity that they introduce in modeling and layout. The capacitance to the shield and the various losses now require much lengthier electromagnetic simulations.

Stacked Inductors At frequencies up to about 5 GHz, inductor values encountered in practice fall in the range of five to several tens of nanohenries. If realized as a single spiral, such inductors occupy a large area and lead to long interconnects between the circuit blocks. This issue can be resolved by exploiting the third dimension, i.e., by stacking spirals. Illustrated in Fig. 7.45, the idea is to place two or more spirals in series, obtaining a higher inductance not only due to the series connection but also as a result of strong mutual coupling. For example, the total inductance in Fig. 7.45 is given by

$$L_{tot} = L_1 + L_2 + 2M. \tag{7.76}$$

Since the lateral dimensions of L_1 and L_2 are much greater than their vertical separation, L_1 and L_2 exhibit almost perfect coupling, i.e., $M \approx L_1 = L_2$ and $L_{tot} \approx 4L_1$. Similarly, n stacked spirals operating in series raise the total inductance by approximately a factor of n^2.

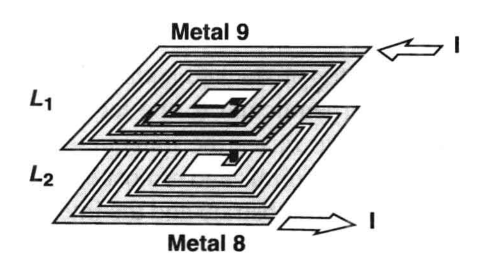


Figure 7.45 *Stacked spirals.*

Example 7.23

The five-turn 4.96-nH inductor obtained from Eq. (7.15) in Section 7.2.3 has an outer dimension of

$$D_{out} = \frac{l_{tot}}{4N} + W + (N-1)(W+S) \tag{7.77}$$

$$= 122\ \mu\text{m}. \tag{7.78}$$

Using Eq. (7.15) for the inductance of one spiral, determine the required outer dimension of a four-turn stacked structure having the same W and S. Assume two spirals are stacked.

(Continues)

Example 7.23 (*Continued*)

Solution:

Each spiral must provide an inductance of 4.96 nH/4 = 1.24 nH. Iteration with $N = 4$, $W = 4\,\mu$m, and $S = 0.5\,\mu$m in Eq. (7.15) yields $l_{tot} \approx 780\,\mu$m and hence $D_{out} = 66.25\,\mu$m. Stacking thus reduces the outer dimension by nearly a factor of 2 in this case.

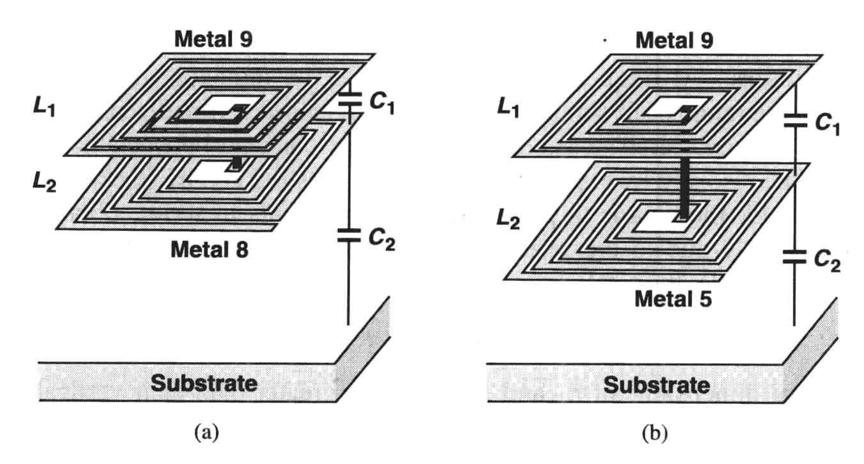


Figure 7.46 *Equivalent capacitance for a stack of (a) metal-9 and metal-8, or (b) metal-9 and metal-5 spirals.*

In reality, the multiplication factor of stacked square inductors is less than n^2 because the legs of one inductor that are perpendicular to the legs of the other provide no mutual coupling. For example, a stack of two raises the inductance by about a factor of 3.5 [6]. The factor is closer to n^2 for octagonal spirals and almost equal to n^2 for circular structures.

In addition to the capacitance to the substrate and the interwinding capacitance, stacked inductors also contain one *between* the spirals [Fig. 7.46(a)].

Example 7.24

In most circuits, one terminal of the inductor(s) is at ac ground. Which terminal of the structure in Fig. 7.46(a) should be grounded?

Solution:

Since L_2 sees a much larger capacitance to the substrate than L_1 does, the terminal of L_2 should be grounded. This is a critical point in the use of stacked inductors.

Using an energy-based analysis similar to that in Section 7.2.4, [6] proves that the equivalent lumped capacitance of the inductor shown in Fig. 7.46(a) is equal to

$$C_{eq} = \frac{4C_1 + C_2}{12}, \tag{7.79}$$

if the free terminal of L_2 is at ac ground.[12] Interestingly, the inter-spiral capacitance has a larger weighting factor than the capacitance to the substrate does. For this reason, if L_2 is moved to lower metal layers [Fig. 7.46(b)], C_{eq} falls even though C_2 rises. Note that the total inductance remains approximately constant so long as the lateral dimensions are much greater than the vertical spacing between L_1 and L_2.

Example 7.25

Compare the equivalent lumped capacitance of single-layer and stacked 4.96-nH inductors studied in Example 7.23. Assume the lower spiral is realized in metal 5 and use the capacitance values shown in Table 7.1.

Table 7.1 *Table of metal capacitances (aF/μm²).*

	Metal 8	Metal 7	Metal 6	Metal 5	Substrate
Metal 9	52	16	12	9.5	4.4
Metal 8		52	24	16	5.4
Metal 7			88	28	6.1
Metal 6				88	7.1
Metal 5					8.6

Solution:

For a single metal-9 layer, the total area is equal to 2000 μm × 4 μm = 8000 μm², yielding a total capacitance of 35.2 fF to the substrate. As suggested by Eq. (7.26), the equivalent lumped capacitance is 1/3 of this value, 11.73 fF. For the stacked structure, each spiral has an area of 780 μm × 4 μm = 3120 μm². Thus, $C_1 = 29.64$ fF and $C_2 = 26.83$ fF, resulting in

$$C_{eq} = 12.1 \text{ fF}. \tag{7.80}$$

The choice of stacking therefore translates to comparable capacitances.[13] If L_2 is moved down to metal 4 or 3, the capacitance of the stacked structure falls more.

For n stacked spirals, it can be proved that

$$C_{eq} = \frac{4\sum_{m=1}^{n-1} C_m + C_{sub}}{3n^2}, \tag{7.81}$$

where C_m denotes each inter-spiral capacitance [6].

12. If the free terminal of L_1 is grounded, the equivalent capacitance is quite larger.
13. We have neglected the fringe components for simplicity.

How does stacking affect the Q? We may surmise that the "resistance-free" coupling, M, among the spirals raises the inductance without increasing the resistance. However, M also exists among the turns of a single, large spiral. More fundamentally, for a given inductance, the total wire's length is relatively constant and independent of how the wire is wound. For example, the single-spiral 4.96-nH inductor studied above has a total length of 2000 μm and the double-spiral stacked structure in Example 7.23, 1560 μm. But, with a more realistic multiplication factor of 3.5 for the inductance of two stacked spirals, the total length grows to about 1800 μm. We now observe that since the top metal layer is typically thicker than the lower layers, stacking tends to *increase* the series resistance and hence decrease the Q. The issue can be remedied by placing two or more lower spirals in *parallel*. Figure 7.47 shows an example where a metal-9 spiral is in series with the parallel combination of metal-6 and metal-5 spirals. Of course, complex current crowding effects at high frequencies require careful electromagnetic field simulations to determine the Q.

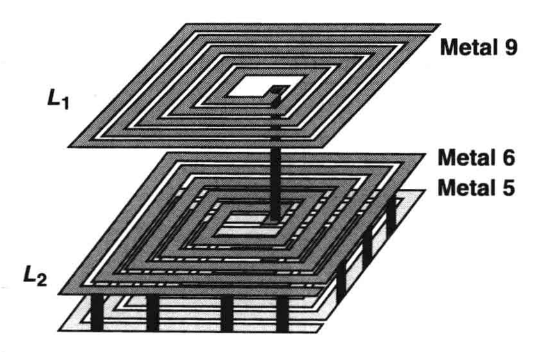


Figure 7.47 *Stacked inductor using two parallel spirals in metal 6 and metal 5.*

7.3 TRANSFORMERS

Integrated transformers can perform a number of useful functions in RF design: (1) impedance matching, (2) feedback or feedforward with positive or negative polarity, (3) single-ended to differential conversion or vice versa, and (4) ac coupling between stages. They are, however, more difficult to model and design than are inductors.

A well-designed transformer must exhibit the following: (1) low series resistance in the primary and secondary windings, (2) high magnetic coupling between the primary and the secondary, (3) low capacitive coupling between the primary and the secondary, and (4) low parasitic capacitances to the substrate. Some of the trade-offs are thus similar to those of inductors.

7.3.1 Transformer Structures

An integrated transformer generally comprises two spiral inductors with strong magnetic coupling. To arrive at "planar" structure, we begin with a symmetric inductor and break it at its point of symmetry (Fig. 7.48). Segments AB and CD now act as mutually-coupled inductors. We consider this structure a 1-to-1 transformer because the primary and the secondary are identical.

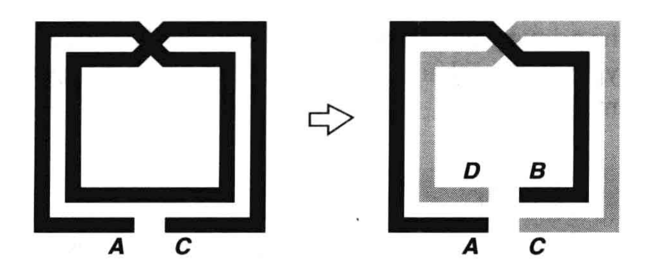


Figure 7.48 *Transformer derived from a symmetric inductor.*

Example 7.26

What is the relationship between the inductance of the symmetric spiral of Fig. 7.48 and the inductances of the resulting transformer?

Solution:

We have

$$L_{AC} = L_{AB} + L_{CD} + 2M, \tag{7.82}$$

where each L refers to the inductance between its end points and M to the mutual coupling between L_{AB} and L_{DC}. Since $L_{AB} = L_{CD}$,

$$L_{AC} = 2L_{AB} + 2M. \tag{7.83}$$

If L_{AC} and M are known, we can determine the inductance of the primary and the secondary.

The transformer structure of Fig. 7.48 suffers from low magnetic coupling, an asymmetric primary, and an asymmetric secondary. To remedy the former, the number of turns can be increased [Fig. 7.49(a)] but at the cost of higher capacitive coupling. To remedy the latter, two symmetric spirals can be embedded as shown in Fig. 7.49(b) but with a slight

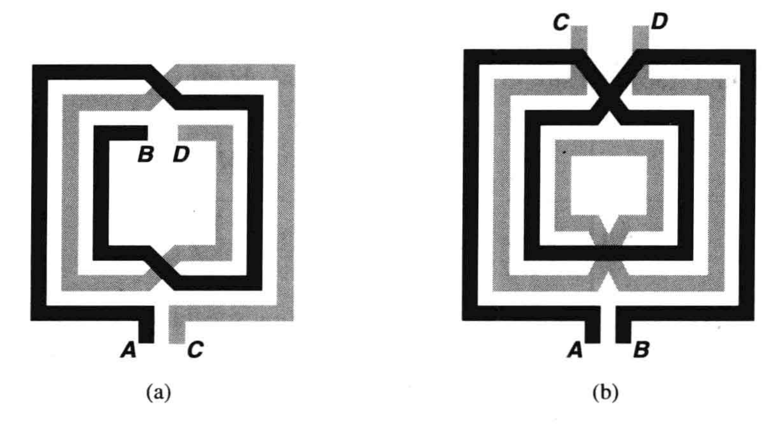


Figure 7.49 *Transformers (a) derived from a three-turn symmetric inductor, (b) formed as two embedded symmetric spirals.*

difference between the primary and secondary inductances. The coupling factor in all of the above structures is typically less than 0.8. We study the consequences of this imperfection in the following example.

Example 7.27

Consider the circuit shown in Fig. 7.50, where C_F models the equivalent lumped capacitance between the primary and the secondary. Determine the transfer function V_{out}/V_{in} and discuss the effect of the sub-unity magnetic coupling factor.

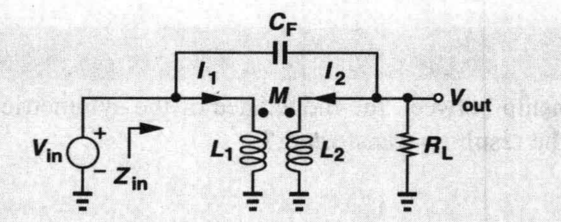


Figure 7.50 *Simple transformer model.*

Solution:

The transformer action gives

$$V_{in} = L_1 s I_1 + M s I_2 \tag{7.84}$$

$$V_{out} = L_2 s I_2 + M s I_1. \tag{7.85}$$

Finding I_1 from Eq. (7.84) and substituting the result in Eq. (7.85), we have

$$I_2 = \frac{V_{out}}{L_2 s} - \frac{M(V_{in} - M s I_2)}{L_1 L_2 s}. \tag{7.86}$$

Also, a KCL at the output node yields

$$(V_{in} - V_{out}) C_F s - I_2 = \frac{V_{out}}{R_L}. \tag{7.87}$$

Replacing I_2 from (7.86) and simplifying the result, we obtain

$$\frac{V_{out}}{V_{in}}(s) = \frac{L_1 L_2 \left(1 - \frac{M^2}{L_1 L_2}\right) C_F s^2 + M}{L_1 L_2 \left(1 - \frac{M^2}{L_1 L_2}\right) C_F s^2 + \frac{L_1 L_2}{R_L}\left(1 - \frac{M^2}{L_1 L_2}\right) s + L_1}. \tag{7.88}$$

It is instructive to examine this transfer function in a few special cases. First, if $C_F = 0$,

$$\frac{V_{out}}{V_{in}} = \frac{M}{\frac{L_1 L_2}{R_L}\left(1 - \frac{M^2}{L_1 L_2}\right) s + L_1}, \tag{7.89}$$

Example 7.27 (*Continued*)

suggesting that, since $k = M/\sqrt{L_1 L_2} < 1$, the transformer exhibits a low-pass response with a real pole located at

$$\omega_p = \frac{-R_L}{L_2\left(1 - \frac{M^2}{L_1 L_2}\right)}. \tag{7.90}$$

For example, if $k = 0.7$, then $\omega_p = -1.96R_L/L_2$. This pole must lie well above the frequency of operation.

Second, if $C_F > 0$ but $M = L_1 = L_2$, then $V_{out}/V_{in} = M/L_1 = 1$ regardless of the values of C_F and R_L. Thus, C_F manifests itself because of the sub-unity k. Since typically $L_1 = L_2 = L$, we can express the poles of Eq. (7.88) as

$$\omega_{p1,2} = \frac{1}{2R_L C_F}\left[-1 \pm \sqrt{1 - \frac{4R_L^2 C_F}{L(1-k^2)}}\right]. \tag{7.91}$$

Equation (7.88) implies that it is beneficial to reduce L_1 and L_2 while k remains constant; as L_1 and L_2 (and $M = k\sqrt{L_1 L_2}$) approach zero,

$$\frac{V_{out}}{V_{in}}(s) \approx \frac{M}{L_1}, \tag{7.92}$$

a frequency-independent quantity equal to k if $L_1 = L_2$. However, reduction of L_1 and L_2 also lowers the input impedance, Z_{in}, in Fig. 7.50. For example, if $C_F = 0$, we have from Eq. (7.54),

$$Z_{in} = L_1 s - \frac{M^2 s^2}{R_L + L_2 s}. \tag{7.93}$$

Thus, the number of primary and secondary turns must be chosen so that Z_{in} is adequately high in the frequency range of interest.

Is it possible to construct planar transformers having a turns ratio greater than unity? Figure 7.51(a) shows an example, where *AB* has approximately one turn and *CD* approximately two. We note, however, that the mutual coupling between *AB* and the inner turn of *CD* is relatively weak due to the smaller diameter of the latter. Figure 7.51(b) depicts another 1-to-2 example with a stronger coupling factor. In practice, the primary and secondary may require a larger number of turns so as to provide a reasonable input impedance.

Figure 7.52 shows two other examples of planar transformers. Here, two asymmetric spirals are interwound to achieve a high coupling factor. The geometry of Fig. 7.52(a) can be viewed as two parallel conductors that are wound into turns. Owing to the difference between their lengths, the primary and secondary exhibit unequal inductances and hence a nonunity turns ratio [16]. The structure of Fig. 7.52(b), on the other hand, provides an exact turns ratio of unity [16].

Transformers can also be implemented as three-dimensional structures. Similar to the stacked inductors studied in Section 7.2.7, a transformer can employ stacked spirals for the

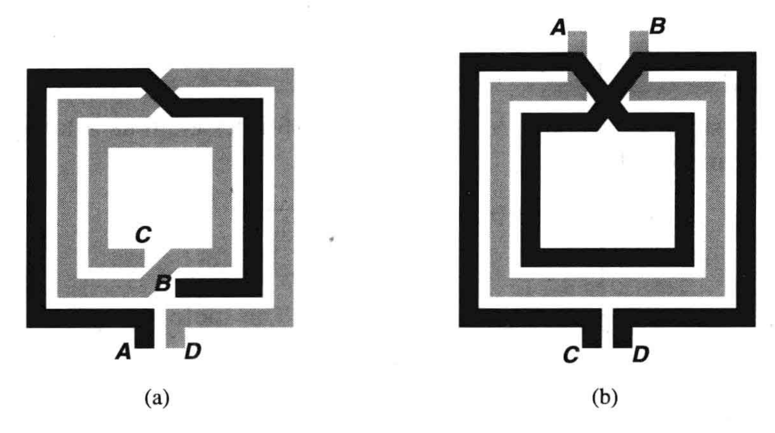


Figure 7.51 *One-to-two transformers (a) derived from a symmetric inductor, (b) formed as two symmetric inductors.*

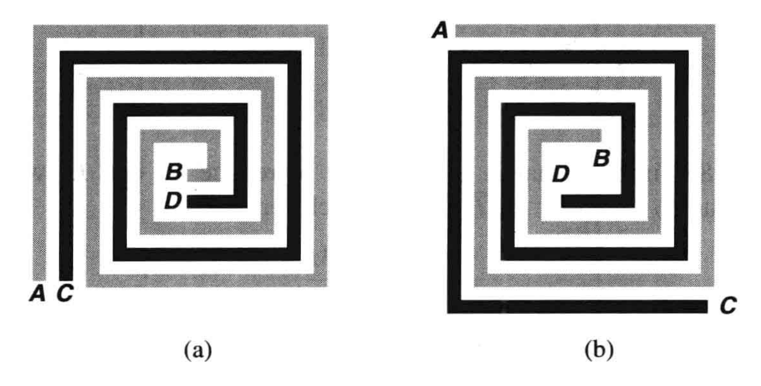


Figure 7.52 *(a) Transformer formed as two wires wound together, (b) alternative version with equal primary and secondary lengths.*

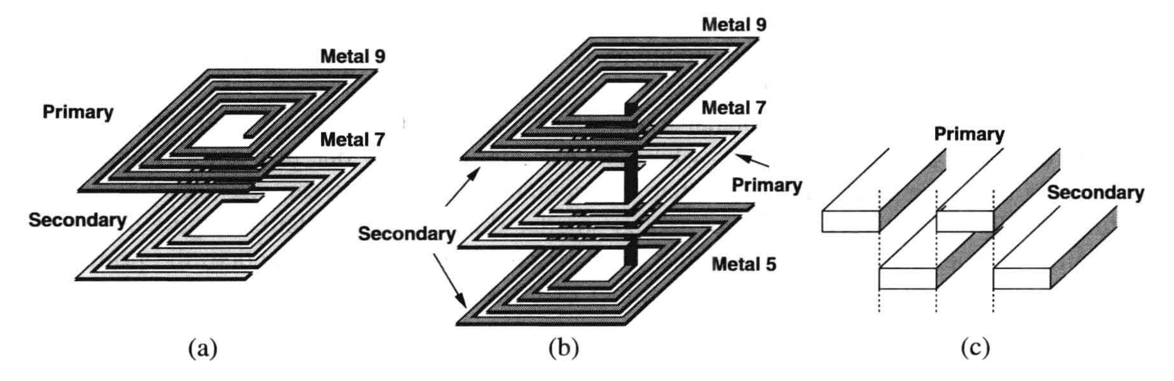


Figure 7.53 *(a) One-to-one stacked transformer, (b) one-to-two transformer, (c) staggering of turns to reduce coupling capacitance.*

primary and the secondary [6]. Figure 7.53(a) shows a 1-to-1 example. It is important to recognize the following attributes: (1) the alignment of the primary and secondary turns results in a slightly higher magnetic coupling factor here than in the planar transformers of Figs. 7.49 and 7.51; (2) unlike the planar structures, the primary and the secondary can be

symmetric *and* identical (except for differences in their capacitances); (3) the overall area occupied by 3D transformers is less than that of their planar counterparts.

Another advantage of stacked transformers is that they can readily provide a turns ratio higher than unity [6]. Illustrated in Fig. 7.53(b), the idea is to incorporate multiple spirals in series to form the primary or the secondary. Thus, a technology having nine metal layers can afford 1-to-8 transformers! As shown in [6], stacked transformers indeed provide significant voltage or current gain at gigahertz frequencies. This "free" gain can be utilized between stages in a chain.

Stacked transformers must, however, deal with two issues. First, the lower spirals suffer from a higher resistance due to the thinner metal layers. Second, the capacitance between the primary and secondary is larger here than in planar transformers (why?). To reduce this capacitance, the primary and secondary turns can be "staggered," thus minimizing their overlap [Fig. 7.53(c)] [6]. But this requires a relatively large spacing between the adjacent turns of each inductor, reducing the inductance.

7.3.2 Effect of Coupling Capacitance

The coupling capacitance between the primary and secondary yields different types of behavior with negative and positive mutual (magnetic) coupling factors. To understand this point, we return to the transfer function in Eq. (7.88) and note that, for $s = j\omega$, the numerator reduces to

$$N(j\omega) = -L_1L_2\left(1 - \frac{M^2}{L_1L_2}\right)C_F\omega^2 + M. \tag{7.94}$$

The first term is always negative, but the polarity of the second term depends on the direction chosen for mutual coupling. Thus, if $M > 0$, then $N(j\omega)$ falls to zero at

$$\omega_z = \sqrt{\frac{M}{L_1L_2\left(1 - \dfrac{M^2}{L_1L_2}\right)C_F}}, \tag{7.95}$$

i.e., the frequency response exhibits a notch at ω_z. On the other hand, if $M < 0$, no such notch exists and the transformer can operate at higher frequencies. We therefore say "noninverting" transformers suffer from a lower speed than do "inverting" transformers [16].

The above phenomenon can also be explained intuitively: the feedforward signal through C_F can cancel the signal coupled from L_1 to L_2. Specifically, the voltage across L_2 in Fig. 7.50 contains two terms, namely, $L_2j\omega I_2$ and $Mj\omega I_1$. If, at some frequency, I_2 is entirely provided by C_F, the former term can *cancel* the latter, yielding a zero output voltage.

7.3.3 Transformer Modeling

An integrated transformer can be viewed as two inductors having magnetic and capacitive coupling. The inductor models described in Section 7.2.6 therefore directly apply here. Figure 7.54 shows an example, where the primary and secondary are represented by the compact inductor model of Fig. 7.35(b), with the mutual coupling M and coupling capacitor C_F added. More details on transformer modeling can be found in [16] and [17].

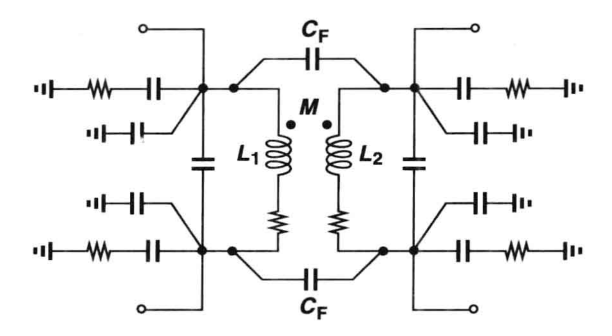


Figure 7.54 *Transformer model.*

Due to the complexity of this model, it is difficult to find the value of each component from measurements or field simulations that provide only *S*- or *Y*-parameters for the entire structure. In practice, some effort is expended on this type of modeling to develop insight into the transformer's limitations, but an accurate representation may require that the designer directly use the *S*- or *Y*-parameters in circuit simulations. Unfortunately, circuit simulators sometimes face convergence difficulties with these parameters.

7.4 TRANSMISSION LINES

Integrated transmission lines (T-lines) are occasionally used in RF design. It is instructive to consider a few examples of T-line applications. Suppose a long wire carries a high-frequency signal from one circuit block to another (Fig. 7.55). The wire suffers from inductance, capacitance, and resistance. If the width of the wire is increased so as to reduce the inductance and series resistance, then the capacitance to the substrate rises. These parasitics may considerably degrade the signal as the frequency exceeds several gigahertz.

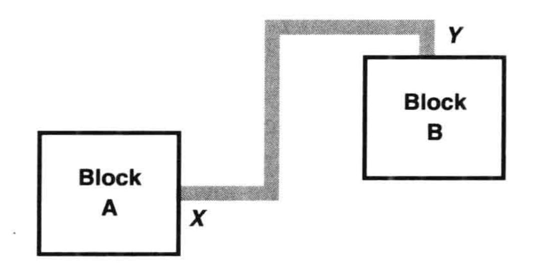


Figure 7.55 *Two circuit blocks connected by a long wire.*

Example 7.28

For the wire shown in Fig. 7.55, we also say the current "return path" is poorly-defined. Explain this attribute and its consequences.

Solution:

In the ideal situation, the signal current flowing through the wire from block *A* to block *B* returns through a ground plane [Fig. 7.56(a)]. In reality, however, due to the wire parasitics

Example 7.28 *(Continued)*

and the nonideal ground connection between the two blocks, some of the signal current flows through the substrate [Fig. 7.56(b)]. The complexity of the return path makes it difficult to accurately predict the behavior of the wire at high frequencies. Also, the coupling to the substrate creates leakage of the signal to other parts of the chip.

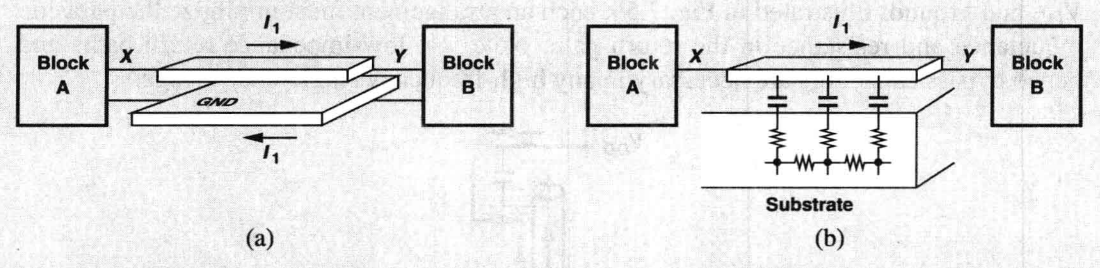


Figure 7.56 *(a) Current return path through a ground plane, (b) poor definition of current return path.*

If the long wire in Fig. 7.55 is replaced with a T-line *and* the input port of block B is modified to match the T-line, then the above issues are alleviated. As illustrated in Fig. 7.57, the line inductance and capacitance no longer degrade the signal, and the T-line ground plane not only provides a low-impedance path for the returning current but minimizes the interaction of the signal with the substrate. The line resistance can also be lowered but with a trade-off (Section 7.4.1).

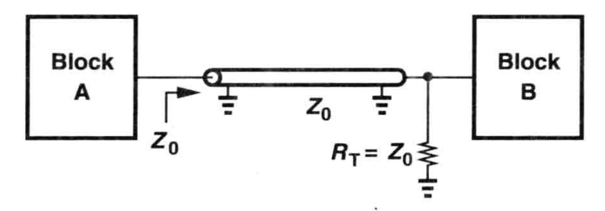


Figure 7.57 *Two circuit blocks connected by a T-line.*

As another example of T-line applications, recall from Chapter 2 that a T-line having a short-circuit termination acts as an inductor if it is much shorter than a wavelength. Thus, T-lines can serve as inductive loads (Fig. 7.58).

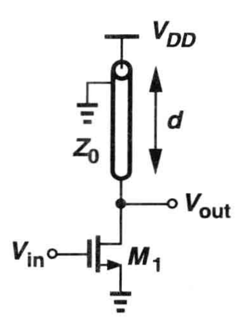


Figure 7.58 *T-line serving as a load inductor.*

Example 7.29

Identify the return path for the signal current that flows through the T-line in Fig. 7.58.

Solution:

Since the signal current reaches the V_{DD} line, a bypass capacitor must be placed between V_{DD} and ground. Illustrated in Fig. 7.59, such an arrangement must minimize the parasitic inductance and resistance in the return path. Note that low-impedance return paths and hence bypass capacitors are necessary in any high-frequency single-ended stage.

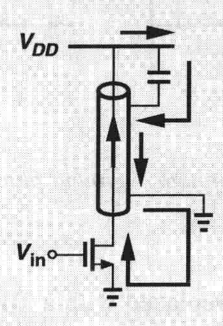


Figure 7.59 *Return path around a T-line in a CS stage.*

How does the Q of T-line inductors compare with that of spiral structures? For frequencies as high as several tens of gigahertz, the latter provide a higher Q because of the mutual coupling among their turns. For higher frequencies, it is expected that the former become superior, but actual measured data supporting this prediction are not available—at least in CMOS technology.

T-lines can also transform impedances. As mentioned in Chapter 2, a line of length d that is terminated with a load impedance of Z_L exhibits an input impedance of

$$Z_{in}(d) = \frac{Z_L + jZ_0 \tan(\beta d)}{Z_0 + jZ_L \tan(\beta d)}, \tag{7.96}$$

where $\beta = 2\pi/\lambda$ and Z_0 is the characteristic impedance of the line. For example, if $d = \lambda/4$, then $Z_{in} = Z_0^2/Z_L$, i.e., a capacitive load can be transformed to an inductive component. Of course, the required quarter-wave length becomes practical in integrated circuits only at millimeter-wave frequencies.

7.4.1 T-Line Structures

Among various T-line structures developed in the field of microwaves, only a few lend themselves to integration. When choosing a geometry, the RF IC designer is concerned with the following parameters: loss, characteristic impedance, velocity, and size.

Before studying T-line structures, let us briefly look at the back end of CMOS processes. As exemplified by Fig. 7.60, a typical process provides a silicided polysilicon layer and about nine metal layers. The high sheet resistance, R_{sh}, of poly (10 to 20 $\Omega/\square$) makes

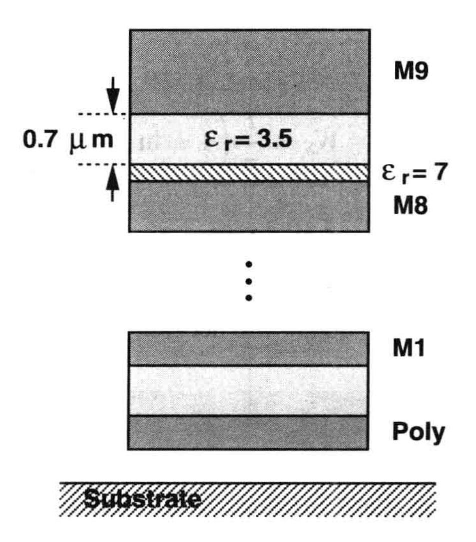


Figure 7.60 *Typical back end of a CMOS process.*

it a poor conductor. Each of the lower metal layers has a thickness of approximately 0.3 μm and an R_{sh} of 60 to 70 mΩ/□. The top layer has a thickness of about 0.7 to 0.8 μm and an R_{sh} of 25 to 30 mΩ/□. Between each two consecutive metal layers lie *two* dielectric layers: a 0.7-μm layer with $\epsilon_r \approx 3.5$ and a 0.1-μm layer with $\epsilon_r \approx 7$.

Microstrip A natural candidate for integrated T-lines is the "microstrip" structure. Depicted in Fig. 7.61, it consists of a signal line realized in the topmost metal layer and a ground plane in a lower metal layer. An important attribute of this topology is that it can have minimal interaction between the signal line and the substrate. This is accomplished if the ground plane is wide enough to contain most of the electric field lines emanating from the signal wire. As a compromise between field confinement and the dimensions of the T-line, we choose $W_G \approx 3W_S$.

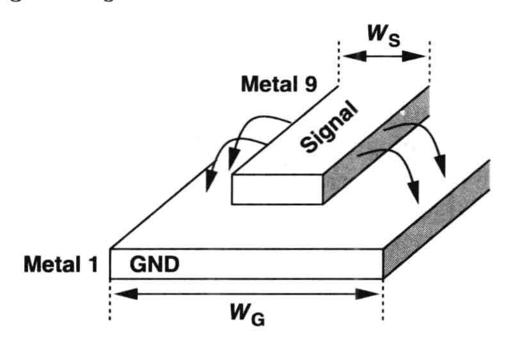


Figure 7.61 *Microstrip structure.*

Numerous equations have been developed in the field of microwaves to express the characteristic impedance of microstrips. For example, if the signal line has a thickness of t and a height of h with respect to the ground plane, then

$$Z_0 = \frac{377}{\sqrt{\epsilon_r}} \frac{h}{W_S} \frac{1}{1 + 1.735\epsilon_r^{-0.0724}(W_e/h)^{-0.836}}, \tag{7.97}$$

where

$$W_e = W_S + \frac{t}{\pi}\left(1 + \ln\frac{2h}{t}\right). \tag{7.98}$$

For example, if $h = 7\ \mu\text{m}$, $t = 0.8\ \mu\text{m}$, $\epsilon_r = 4$, and $W_S = 4\ \mu\text{m}$, then $Z_0 \approx 86\ \Omega$. Unfortunately, these equations suffer from errors as large as 10%. In practice, electromagnetic field simulations including the back end details are necessary to compute Z_0.

Example 7.30

A short microstrip is used as an inductor resonating with the transistor capacitances in a circuit. Determine the error in the resonance frequency, ω_{res}, if the line's characteristic impedance has a 10% error.

Solution:

From Eq. (7.96), a T-line with $Z_L = 0$ and $2\pi d \ll \lambda$ provides an input impedance of

$$Z_{in} = jZ_0 \tan(\beta d) \tag{7.99}$$

$$\approx jZ_0\left(2\pi\frac{d}{\lambda}\right) \tag{7.100}$$

$$\approx j\omega\frac{Z_0 d}{v}, \tag{7.101}$$

i.e., an inductance of $L_{eq} = Z_0 d/v = L_u d$, where v denotes the wave velocity and L_u the inductance per unit length. Since ω_{res} is inversely proportional to $\sqrt{L_{eq}}$, a 10% error in L_{eq} translates to about a 5% error in ω_{res}.

The loss of microstrips arises from the resistance of both the signal line and the ground plane. In modern CMOS technologies, metal 1 is in fact thinner than the higher layers, introducing a ground plane loss comparable to the signal line loss.

The loss of a T-line manifests itself as signal attenuation (or bandwidth reduction) if the line simply connects two blocks. With a typical loss of less than 0.5 dB/mm at frequencies of several tens of gigahertz, a microstrip serves this purpose well. On the other hand, if a T-line acts as an inductive load whose Q is critical, then a much lower loss is required. We can readily relate the loss and the Q. Suppose a T-line of unit length exhibits a series resistance of R_u. As shown in Fig. 7.62,

$$\frac{V_{out}}{V_{in}} \approx \frac{R_L}{R_S + R_u + R_L} \tag{7.102}$$

$$\approx \frac{Z_0}{2Z_0 + R_u}. \tag{7.103}$$

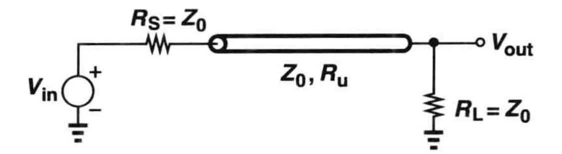


Figure 7.62 *Lossy transmission line.*

We find the difference between this result and the ideal value and then normalize to 1/2:

$$\text{Loss} \approx \frac{R_u}{2Z_0 + R_u} \tag{7.104}$$

$$\approx \frac{R_u}{2Z_0}, \tag{7.105}$$

if $R_u \ll 2Z_0$. Note that this value is expressed in decibels as $20\log(1 - \text{Loss})$ and the result is *negative*. A T-line of unit length has a Q of

$$Q = \frac{L_u\omega}{R_u} \tag{7.106}$$

$$= \frac{L_u\omega}{2Z_0 \cdot \text{Loss}}. \tag{7.107}$$

Example 7.31

Consider a microstrip line 1000 μm long with $Z_0 = 100\ \Omega$ and $L = 1$ nH. If the signal line is 4 μm wide and has a sheet resistance of 25 mΩ/□, determine the loss and the Q at 5 GHz. Neglect skin effect and the loss of the ground plane.

Solution:

The low-frequency resistance of the signal line is equal to 6.25 Ω, yielding from Eq. (7.104) a loss of $0.031 \equiv -0.276$ dB. The Q is obtained from (7.107) as

$$Q = 5.03. \tag{7.108}$$

In order to reduce the loss of a microstrip, the width of the signal line can be increased (requiring a proportional increase in the width of the ground plane). But such an increase (1) reduces the inductance per unit length (as if multiple signal lines were placed in parallel), and (2) raises the capacitance to the ground plane. Both effects translate to a lower characteristic impedance, $Z_0 = \sqrt{L_u/C_u}$. For example, doubling the signal line width roughly halves Z_0.[14] Equation (7.97) also reveals this rough dependence.

The reduction of the characteristic impedance as a result of widening the signal line does make circuit design more difficult. As noted in Fig. 7.57, a properly-terminated T-line

14. Doubling the width does not reduce L_u by a factor of 2 because placing two *coupled* wires in parallel does not halve the inductance.

loads the driving stage (block A) with a resistance of Z_0. Thus, as Z_0 decreases, so does the gain of block A. In other words, it is the *product* of the gain of block A and the inverse loss of the T-line that must be maximized, dictating that the circuit and the line be designed as a single entity.

The resistance of microstrips can also be reduced by stacking metal layers. Illustrated in Fig. 7.63, such a geometry alleviates the trade-off between the loss and the characteristic impedance. Also, stacking allows a narrower footprint for the T-line, thus simplifying the routing and the layout.

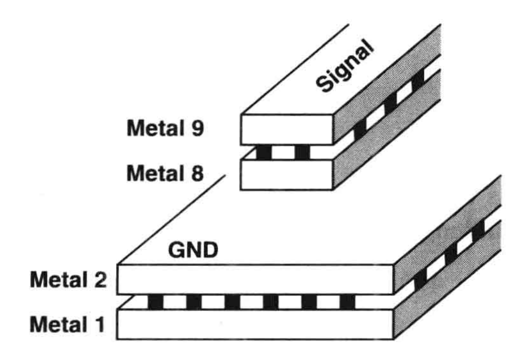


Figure 7.63 *Microstrip using parallel metal layers for lower loss.*

Example 7.32

Transmission lines used to transform impedances are prohibitively long for frequencies up to a few tens of gigahertz. However, the relationship $v = 1/\sqrt{L_u C_u}$ suggests that, if C_u is raised, then the wave velocity can be reduced and so can $\lambda = v/f$. Explain the practicality of this idea.

Solution:

The issue is that a higher C_u results in a lower Z_0. Thus, the line can be shorter, but it demands a greater drive capability. Moreover, impedance transformation becomes more difficult. For example, suppose a $\lambda/4$ line is used to *raise* Z_L to Z_0^2/Z_L. This is possible only if $Z_0 > Z_L$.

Coplanar Lines Another candidate for integrated T-lines is the "coplanar" structure. Shown in Fig. 7.64, this geometry realizes both the signal and the ground lines in *one plane*, e.g., in metal 9. The characteristic impedance of coplanar lines can be higher than that of microstrips because (1) the thickness of the signal and ground lines in Fig. 7.64 is quite small, leading to a small capacitance between them, and (2) the spacing between the two lines can be large, further decreasing the capacitance. Of course, as S becomes comparable with h, more of the electric field lines emanating from the signal wire terminate on the substrate, producing a higher loss. Also, the signal line can be surrounded by ground lines on both sides. The characteristics of coplanar lines are usually obtained by electromagnetic field simulations.

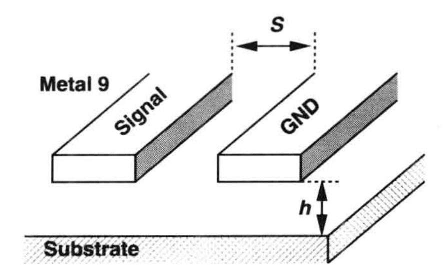


Figure 7.64 *Coplanar structure.*

The loss reduction techniques described above for microstrips can also be applied to coplanar lines, entailing similar trade-offs. However, coplanar lines have a larger footprint because of their lateral spread, making layout more difficult.

Stripline The "stripline" consists of a signal line *surrounded* by ground planes, thus producing little field leakage to the environment. As an example, a metal-5 signal line can be surrounded by metal-1 and metal-9 planes and vias connecting the two planes (Fig. 7.65). If the vias are spaced closely, the signal line remains shielded in all four directions.

The stripline exhibits a smaller characteristic impedance than microstrip and coplanar structures do. It is therefore used only where field confinement is essential.

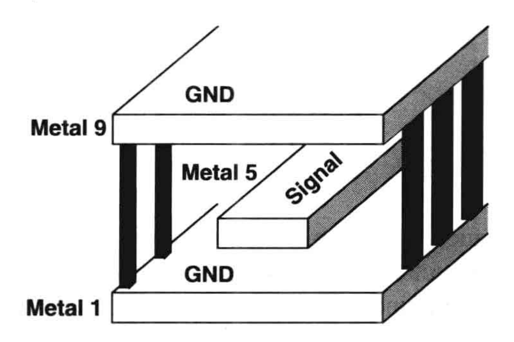


Figure 7.65 *Stripline structure.*

7.5 VARACTORS

As described in Chapter 8, "varactors" are an essential component of LC VCOs. Varactors also occasionally serve to tune the resonance frequency of narrowband amplifiers.

A varactor is a voltage-dependent capacitor. Two attributes of varactors become critical in oscillator design: (1) the capacitance range, i.e., the ratio of the maximum and minimum capacitances that the varactor can provide, and (2) the quality factor of the varactor, which is limited by the parasitic series resistances within the structure. Interestingly, these two parameters trade with each other in some cases.

In older generations of RF ICs, varactors were realized as reverse-biased *pn* junctions. Illustrated in Fig. 7.66(a) is one example where the *p*-substrate forms the anode and the n^+ contact, the cathode. (The p^+ contact provides a low-resistance connection to

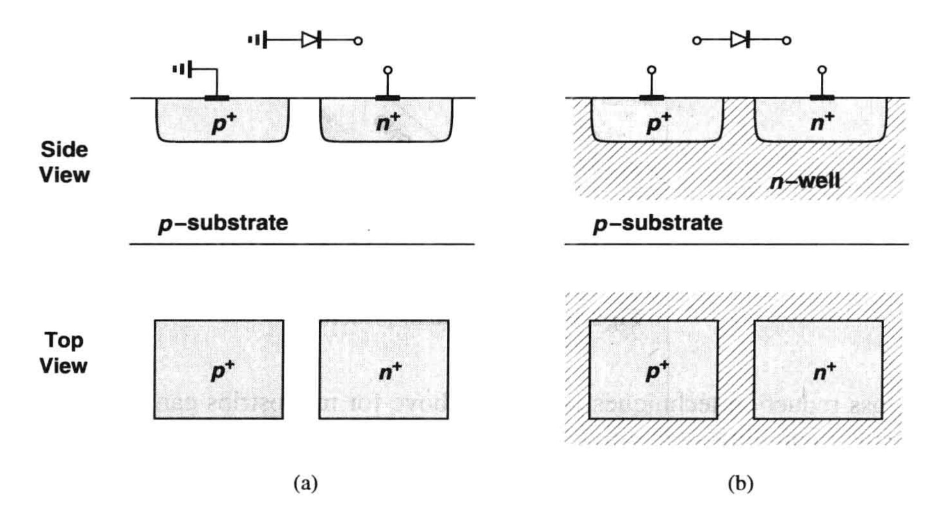


Figure 7.66 *PN junction varactor with (a) one terminal grounded, (b) both terminals floating.*

the substrate.) In this case, the anode is "hard-wired" to ground, limiting the design flexibility. A "floating" *pn* junction can be constructed as shown in Fig. 7.66(b), with an *n*-well isolating the diode from the substrate and acting as the cathode.

Let us examine the capacitance range and Q of *pn* junctions. At a reverse bias of V_D, the junction capacitance, C_j, is given by

$$C_j = \frac{C_{j0}}{\left(1 + \dfrac{V_D}{V_0}\right)^m}, \tag{7.109}$$

where C_{j0} is the capacitance at zero bias, V_0 the built-in potential, and m an exponent around 0.3 in integrated structures. We recognize the weak dependence of C_j upon V_D. Since $V_0 \approx 0.7$ to 0.8 V and since V_D is constrained to less than 1 V by today's supply voltages, the term $1 + V_D/V_0$ varies between approximately 1 and 2. Furthermore, an m of about 0.3 weakens this variation, resulting in a capacitance range, $C_{j,max}/C_{j,min}$, of roughly 1.23. In practice, we may allow the varactor to experience some forward bias (0.2 to 0.3 V), thus obtaining a somewhat larger range.

The Q of a *pn*-junction varactor is given by the total series resistance of the structure. In the floating diode of Fig. 7.66(b), this resistance is primarily due to the *n*-well and can be minimized by selecting minimum spacing between the n^+ and p^+ contacts. Moreover, as shown in Fig. 7.67, each p^+ region can be surrounded by an n^+ ring to lower the resistance in two dimensions.

Unlike inductors, transformers, and T-lines, varactors are quite difficult to simulate and model, especially for Q calculations. Consider the displacement current flow depicted in Fig. 7.68(a). Due to the two-dimensional nature of the flow, it is difficult to determine or compute the equivalent series resistance of the structure. This issue arises partly because the sheet resistance of the *n*-well is typically measured by the foundry for contacts having a spacing greater than the depth of the *n*-well [Fig. 7.68(b)]. Since the current path in this case is different from that in Fig. 7.68(a), the *n*-well sheet resistance cannot be directly applied

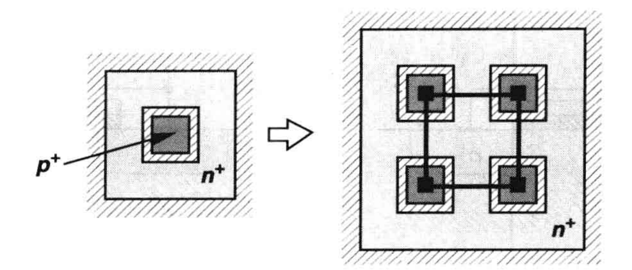


Figure 7.67 *Use of an n^+ ring to reduce varactor resistance.*

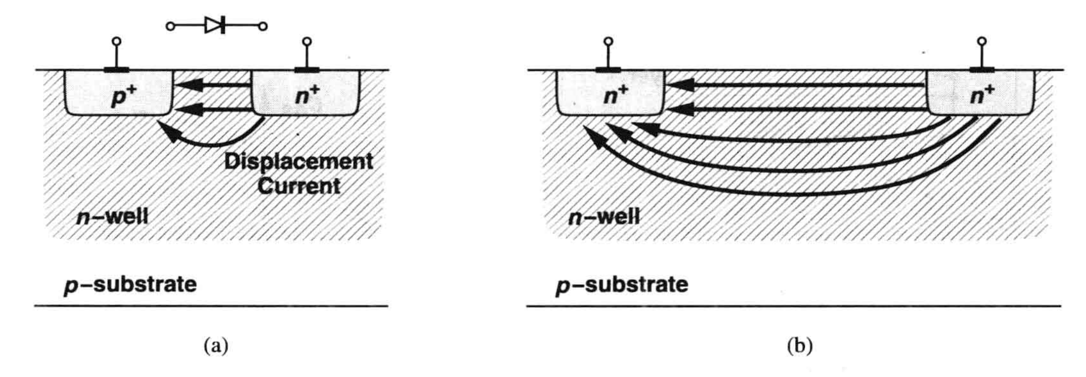


Figure 7.68 *Current distribution in a (a) varactor, (b) typical test structure.*

to the calculation of the varactor series resistance. For these reasons, the Q of varactors is usually obtained by measurement on fabricated structures.[15]

In modern RF IC design, MOS varactors have supplanted their *pn*-junction counterparts. A regular MOSFET exhibits a voltage-dependent gate capacitance (Fig. 7.69), but the nonmonotonic behavior limits the design flexibility. For example, a voltage-controlled oscillator (VCO) employing such a varactor would generate an output frequency that rises *and* falls as (the average) V_{GS} goes from negative to positive values. This nonmonotonic frequency tuning behavior becomes problematic in phase-locked loop design (Chapter 9).

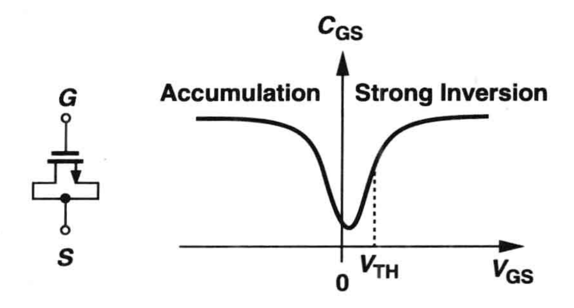


Figure 7.69 *Variation of gate capacitance with V_{GS}.*

15. Of course, semiconductor device simulators can be used here if the doping levels and the junction depths are known.

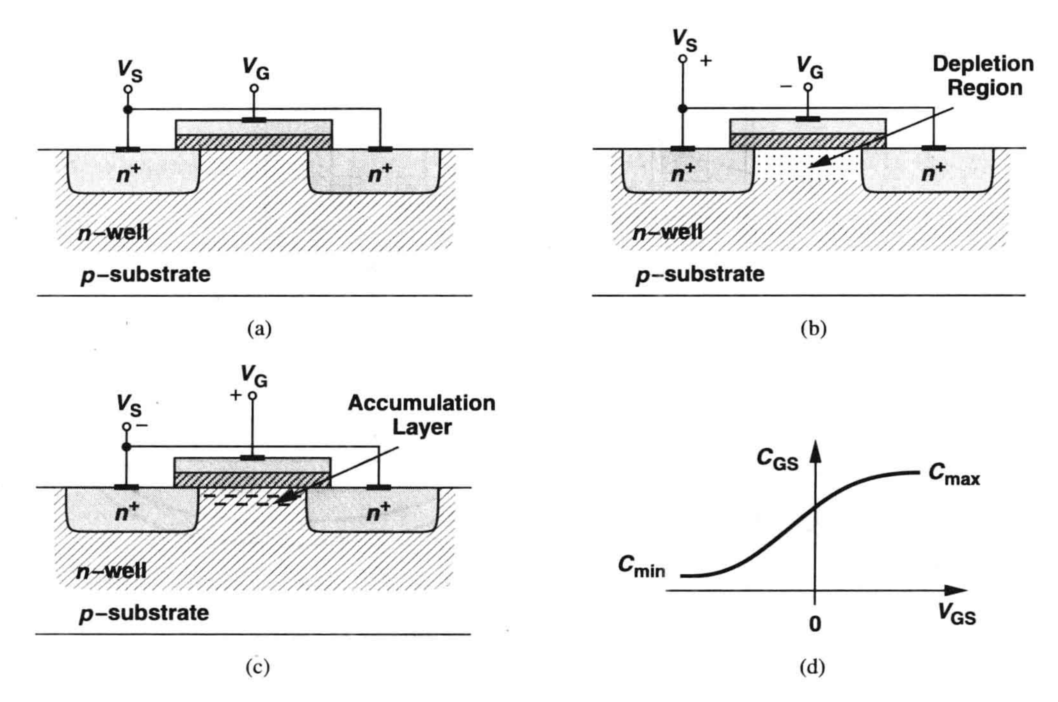


Figure 7.70 *(a) MOS varactor, (b) operation with negative gate-source voltage, (c) operation with positive gate-source voltage, (d) resulting C/V characteristic.*

A simple modification of the MOS device avoids the above issues. Called an "accumulation-mode MOS varactor" and shown in Fig. 7.70(a), this structure is obtained by placing an NMOS transistor inside an n-well. If $V_G < V_S$, then the electrons in the n-well are repelled from the silicon/oxide interface and a depletion region is formed [Fig. 7.70(b)]. Under this condition, the equivalent capacitance is given by the series combination of the oxide and depletion capacitances. As V_G exceeds V_S, the interface attracts electrons from the n^+ source/drain terminals, creating a channel [Fig. 7.70(c)]. The overall capacitance therefore rises to that of the oxide, behaving as shown in Fig. 7.70(d). (Since the material under the gate is n-type silicon, the concept of strong inversion does not apply here.)

The C/V characteristic of MOS varactors has scaled well with CMOS technology generations, approaching its saturated levels of C_{max} and C_{min} for $V_{GS} \approx \pm 0.5$ V in 65-nm devices. These varactors therefore operate with low supply voltages better than their pn-junction counterparts.

Another advantage of accumulation-mode MOS varactors is that, unlike pn junctions, they can tolerate both positive and negative voltages. In fact, the characteristic of Fig. 7.70(d) suggests that MOS varactors *should* operate with positive and negative biases so as to provide maximum tuning range. We pursue this point in VCO design in Chapter 8.

Circuit simulations must somehow incorporate the varactor C/V characteristic of Fig. 7.70(d). In practice, this characteristic is measured on fabricated devices and represented by a table of discrete values. Such a table, however, may introduce discontinuities in the *derivatives* of the characteristic, creating undesirable artifacts (e.g., a high noise floor) in simulations. It is therefore desirable to approximate the C/V plot by a well-behaved

function. The hyperbolic tangent proves useful here for both its saturating behavior and its continuous derivatives. Noting that $\tanh(\pm\infty) = \pm 1$, we approximate the characteristic of Fig. 7.70(d) by

$$C_{var}(V_{GS}) = \frac{C_{max} - C_{min}}{2} \tanh\left(a + \frac{V_{GS}}{V_0}\right) + \frac{C_{max} + C_{min}}{2}. \tag{7.110}$$

Here, a and V_0 allow fitting for the intercept and the slope, respectively, and C_{min} and C_{max} include the gate-drain and gate-source overlap capacitance.

The above varactor model translates to different characteristics in different circuit simulators! For example, HSPICE predicts a narrower oscillator tuning range than Cadence does. Simulation tools that analyze circuits in terms of voltages and currents (e.g., HSPICE) interpret the nonlinear capacitance equation correctly. On the other hand, programs that represent the behavior of capacitors by *charge equations* (e.g., Cadence's Spectre) require that the model be transformed to a Q/V relationship. To this end, we recall the general definition of capacitance from $dQ = C(V)dV$ and write

$$Q_{var} = \int C_{var} dV_{GS} \tag{7.111}$$

$$= \frac{C_{max} - C_{min}}{2} V_0 \ln\left[\cosh\left(a + \frac{V_{GS}}{V_0}\right)\right] + \frac{C_{max} + C_{min}}{2} V_{GS}. \tag{7.112}$$

In other words, the varactor is represented as a two-terminal device whose charge and voltage are related by Eq. (7.112). The simulation tool then computes the current flowing through the varactor as

$$I_{var} = \frac{dQ_{var}}{dt}. \tag{7.113}$$

The Q of MOS varactors is determined by the resistance between the source and drain terminals.[16] As shown in Fig. 7.71(a), this resistance and the capacitance are distributed from the source to the drain and can be approximated by the lumped model depicted in Fig. 7.71(b).

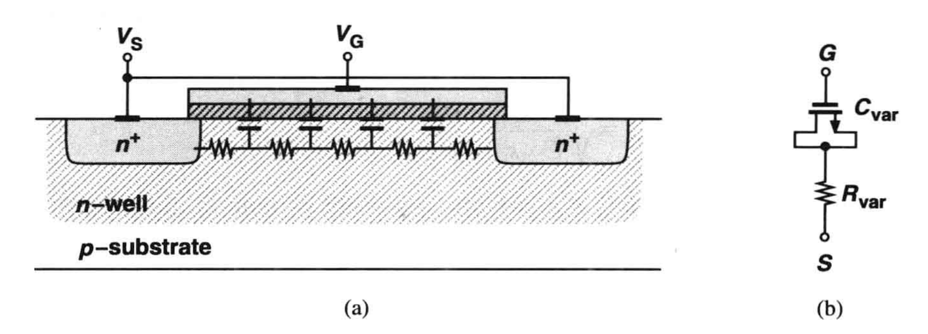


Figure 7.71 *(a) Effect of distributed resistance in a varactor, (b) lumped model.*

16. We assume that the gate resistance is minimized by proper layout.

Example 7.33

Determine the equivalent resistance and capacitance values in the lumped model of Fig. 7.71(b).

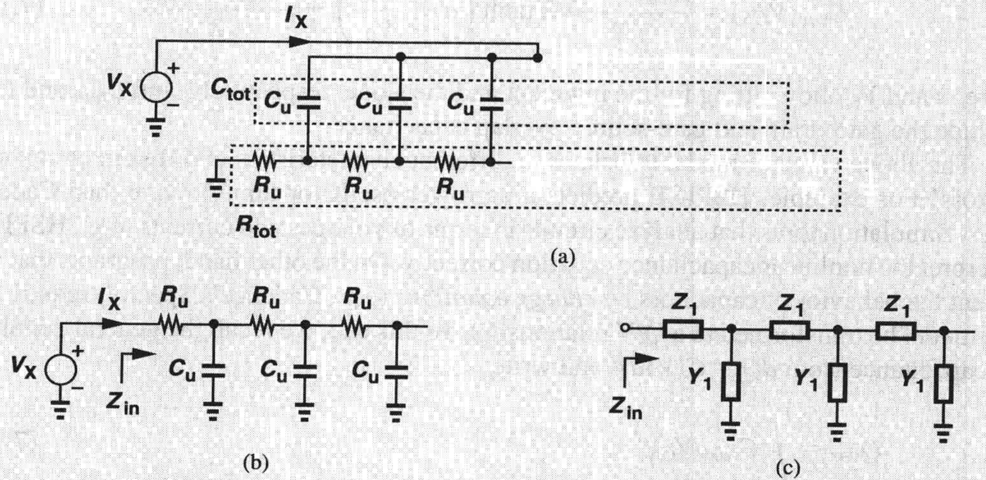


Figure 7.72 *(a) Distributed model of a varactor, (b) equivalent circuit for half of the structure, (c) canonical T-line structure.*

Solution:

Let us first consider only one-half of the structure as shown in Fig. 7.72(a). Here, the unit capacitances add up to the total distributed capacitance, C_{tot}, and the unit resistances to the total distributed resistance, R_{tot}. We turn the circuit upside down, arriving at the more familiar topology illustrated in Fig. 7.72(b). The circuit now resembles a transmission line consisting of series resistances and parallel capacitances. For the general T-line shown in Fig. 7.72(c), it can be proved that the input impedance, Z_{in}, is given by [18]

$$Z_{in} = \sqrt{\frac{Z_1}{Y_1}} \frac{1}{\tanh(\sqrt{Z_1 Y_1} d)}, \tag{7.114}$$

where Z_1 and Y_1 are specified for unit length and d is the length of the line. From Fig. 7.72(b), $Z_1 d = R_{tot}$ and $Y_1 d = C_{tot} s$; thus,

$$Z_{in} = \sqrt{\frac{R_{tot}}{C_{tot} s}} \frac{1}{\tanh(\sqrt{R_{tot} C_{tot} s/4})}. \tag{7.115}$$

At frequencies well below $1/(R_{tot}C_{tot}/4)$, the argument of tanh is much less than unity, allowing the approximation,

$$\tanh \epsilon \approx \epsilon - \frac{\epsilon^3}{3} \tag{7.116}$$

$$\approx \frac{\epsilon}{1 + \frac{\epsilon^2}{3}}. \tag{7.117}$$

Example 7.33 *(Continued)*

It follows that

$$Z_{in} \approx \frac{1}{C_{tot}s/2} + \frac{R_{tot}/2}{3}. \tag{7.118}$$

That is, the lumped model of half of the structure consists of its distributed capacitance in series with one-third of its distributed resistance. Accounting for the gray half in Fig. 7.72(b), we obtain

$$Z_{in,tot} \approx \frac{1}{C_{tot}s} + \frac{R_{tot}}{12}. \tag{7.119}$$

The principal difficulty in computing the Q of MOS varactors (placed inside an n-well) is that the resistance between the source and drain cannot be directly computed from the MOS transistor characteristics. As with pn junctions, the Q of MOS varactors is usually obtained from experimental measurements.

How does the Q of MOS varactors vary with the capacitance? In the characteristic of Fig. 7.70(d), as we begin from C_{min}, the capacitance is small and the resistance somewhat large (that of n-well). On the other hand, as we approach C_{max}, the capacitance rises and the resistance falls. Consequently, equation $Q = 1/(RC\omega)$ suggests that the Q may remain relatively constant. In practice, however, the Q drops as C_{GS} goes from C_{min} to C_{max} (Fig. 7.73), indicating that the relative rise in the capacitance is greater than the relative fall in the resistance.

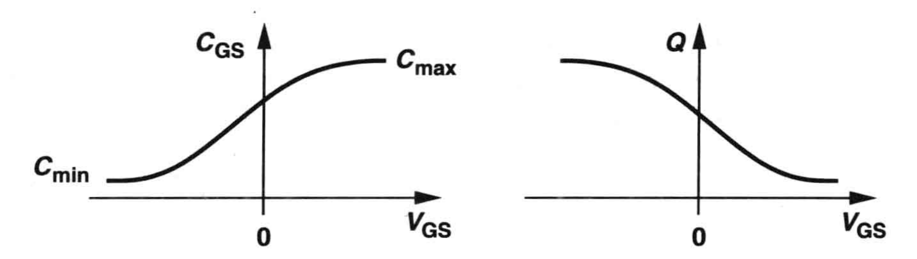


Figure 7.73 *Variation of varactor Q with capacitance.*

As explained in Chapter 8, it is desirable to maximize the Q of varactors for oscillator design. From our foregoing study of MOS varactors, we conclude that the device length (the distance between the source and drain) must be minimized. Unfortunately, for a minimum channel length, the overlap capacitance between the gate and source/drain terminals becomes a substantial fraction of the overall capacitance, limiting the capacitance range. As illustrated in Fig. 7.74, the overlap capacitance (which is relatively voltage-independent) shifts the C/V characteristic up, yielding a ratio of $(C_{max} + 2WC_{ov})/(C_{min} + 2WC_{ov})$, where C_{max} and C_{min} denote the "intrinsic" values, i.e., those without the overlap effect. For a minimum channel length, $2WC_{ov}$ may even be larger than C_{min}, thus reducing the capacitance ratio considerably.

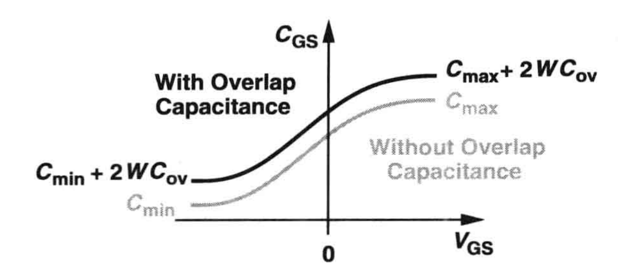


Figure 7.74 *Effect of overlap capacitance on varactor capacitance range.*

Example 7.34

A MOS varactor realized in 65-nm technology has an effective length of 50 nm and a C_{ov} of 0.09 fF/μm. If $C_{ox} = 17\,\text{fF}/\mu\text{m}^2$, determine the largest capacitance range that the varactor can provide.

Solution:

Assuming a width of 1 μm for the device, we have $2WC_{ov} = 0.18\,\text{fF}$ and a gate oxide capacitance of $17\,\text{fF}/\mu\text{m}^2 \times 1\,\mu\text{m} \times 50\,\text{nm} = 0.85\,\text{fF}$. Thus, the minimum capacitance is 0.18 fF (if the series combination of the oxide and depletion capacitances is neglected), and the maximum capacitance reaches $0.85\,\text{fF} + 0.18\,\text{fF} = 1.03\,\text{fF}$. The largest possible capacitance ratio is therefore equal to 5.72. In practice, the series combination of the oxide and depletion capacitances is comparable to $2WC_{ov}$, reducing this ratio to about 2.5.

In order to achieve a larger capacitance range, the length of MOS varactors can be increased. In the above example, if the effective channel length grows to 100 nm, then the capacitance ratio reaches $(1.7\,\text{fF} + 0.18\,\text{fF})/(0.18\,\text{fF}) = 10.4$. However, the larger source-drain resistance results in a lower Q. Since the maximum capacitance goes from 1.03 fF to 1.88 fF and since the channel resistance is doubled, the $Q\ [= 1/(RC\omega)]$ falls by a factor of 3.65. In other words, an m-fold increase in the channel length translates to roughly an m^2-fold drop in the Q.

The trade-off between the capacitance range and Q of varactors ultimately leads to another between the tuning range and phase noise of LC VCOs. We study this issue in Chapter 8. At frequencies up to about 10 GHz, a channel length of twice the minimum may be chosen so as to widen the capacitance range while retaining a varactor Q much larger than the inductor Q.

7.6 CONSTANT CAPACITORS

RF circuits employ constant capacitors for various purposes, e.g., (1) to adjust the resonance frequency of LC tanks, (2) to provide ac coupling between stages, or (3) to bypass the supply rail to ground. The critical parameters of capacitors used in RF ICs include the

capacitance density (the amount of capacitance per unit area on the chip), the parasitic capacitances, and the Q.

7.6.1 MOS Capacitors

MOSFETs configured as capacitors offer the highest density in integrated circuits because C_{ox} is larger than other capacitances in CMOS processes. However, the use of MOS capacitors entails two issues. First, to provide the maximum capacitance, the device requires a V_{GS} higher than the threshold voltage (Fig. 7.69). A similar "bias" requirement applies to MOS varactors if they are to provide maximum capacitance. Second, the channel resistance limits the Q of MOS capacitors at high frequencies. From Eq. (7.119), we note that the channel resistance is divided by 12 in the lumped model, yielding

$$Q = \frac{12}{R_{tot}C_{tot}\omega}. \tag{7.120}$$

Both of the above issues make MOS capacitors a poor choice for interstage coupling. Depicted in Fig. 7.75(a) is an example, wherein M_3 sustains a bias gate-source voltage approximately equal to $V_{DD} - V_{GS2}$ (why?). With typical values of $V_{DD} = 1$ V and $V_{GS2} = 0.5$ V, M_3 suffers from a small overdrive voltage and hence a high channel resistance. Moreover, the nonlinearity of the capacitance of M_3 may manifest itself if the circuit senses large interferers. For these reasons, MOS capacitors rarely serve as coupling devices.

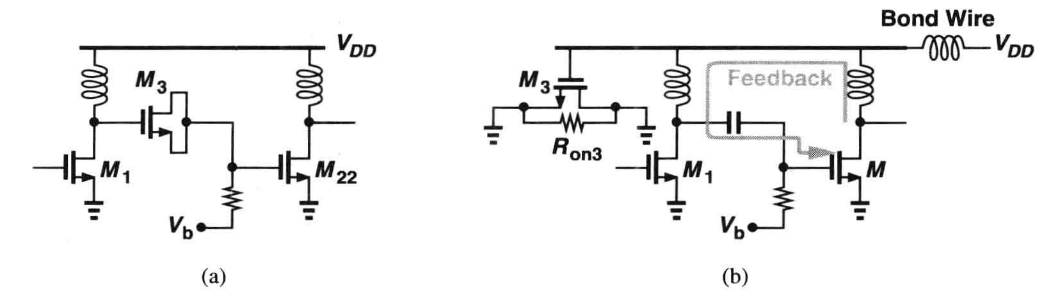


Figure 7.75 *MOS capacitor used as (a) coupling device (b) bypass component.*

One application of MOS capacitors is in supply bypass. As illustrated in Fig. 7.75(b), the supply line may include significant bond wire inductance, allowing *feedback* from the second stage to the first at high frequencies. The bypass capacitor, M_3, creates a low impedance between the supply and the ground, suppressing the feedback. In this case, the Q of M_3 is still important: if the equivalent series resistance of the device becomes comparable with the reactance of its capacitance, then the bypass impedance may not be low enough to suppress the feedback.

It is important to note that typical MOS models fail to include the channel resistance, R_{on}, if the source and the drain are shorted. As illustrated in Fig. 7.75(b) for M_3, R_{on3} is represented as a single lumped component between the two terminals and simply "shorted out" by circuit simulators. For this reason, the designer must compute R_{on} from I/V characteristics, divide it by 12, and insert the result in series with the MOS capacitor.

Example 7.35

A MOS capacitor can be constructed as a single transistor of length L [Fig. 7.76(a)] or N transistors in parallel, each of length L/N. Compare the Q's of the two structures. For simplicity, assume the effective channel lengths are equal to L and L/N, respectively.

Solution:

The structure of Fig. 7.76(a) exhibits a channel resistance of

$$R_{on,a} = \frac{1}{\mu_n C_{ox} \frac{W}{L}(V_{GS} - V_{TH})}, \tag{7.121}$$

and each finger in Fig. 7.76(b) a channel resistance of

$$R_{on,u} = \frac{1}{\mu_n C_{ox} \frac{W}{L/N}(V_{GS} - V_{TH})}. \tag{7.122}$$

Since N fingers appear in parallel, $R_{on,b} = R_{on,u}/N = R_{on,a}/N^2$. That is, the decomposition of the device into N parallel fingers reduces the resistance by a factor of N^2.

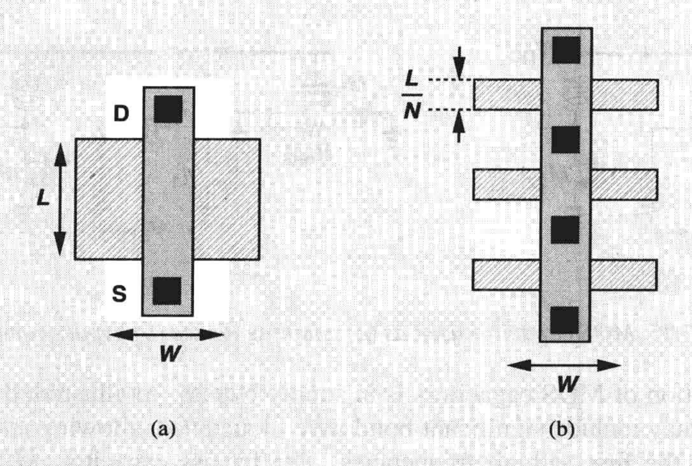


Figure 7.76 *MOS capacitor realized as (a) one long finger, (b) multiple short fingers.*

For frequencies up to a few tens of gigahertz, the above decomposition can yield reasonable Q's (e.g., 5 to 10), allowing the use of MOS capacitors for supply bypass.

The reader is cautioned that very large MOS capacitors suffer from significant gate leakage current, especially with a V_{GS} as high as V_{DD}. This current manifests itself if the system must enter a low-power (standby) mode: the leakage persists as long as V_{DD} is applied, draining the battery.

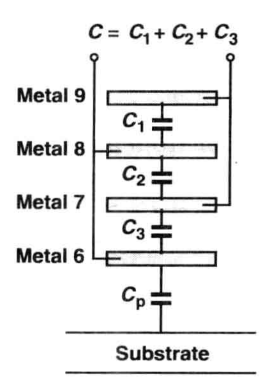


Figure 7.77 *Parallel-plate capacitor.*

7.6.2 Metal-Plate Capacitors

If the Q or linearity of MOS capacitors is inadequate, metal-plate capacitors can be used instead. The "parallel-plate" structure employs planes in different metal layers as shown in Fig. 7.77. For maximum capacitance density, all metal layers (and even the poly layer) can be utilized.

Example 7.36

Show the actual connections necessary among the metal layers shown in Fig. 7.77.

Solution:

The even-numbered metal layers must be tied to one another and so must the odd-numbered layers. As shown in Figure 7.78, these connections are made by vias. In practice, a row of vias (into the page) is necessary to connect the layers so as to obtain a small series resistance.

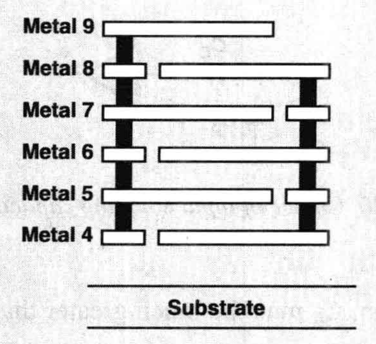


Figure 7.78 *Detailed realization of parallel-plate capacitor.*

The Q and linearity of well-designed parallel-plate capacitors are typically so high that they need not be taken into account. However, even with all metal layers and a poly

layer, parallel-plate structures achieve less capacitance density than MOSFETs do. For example, with nine metal layers in 65-nm technology, the former provides a density of about 1.4 fF/μm^2 and the latter, 17 fF/μm^2.

Parallel-plate geometries also suffer from a parasitic capacitance to the substrate. As illustrated in Fig. 7.79, the capacitance between the lowest plate and the substrate, C_p, divided by the desired capacitance, $C_{AB} = C_1 + \cdots + C_9$, represents the severity of this parasitic. In a typical process, this value reaches 10%, leading to serious difficulties in circuit design.

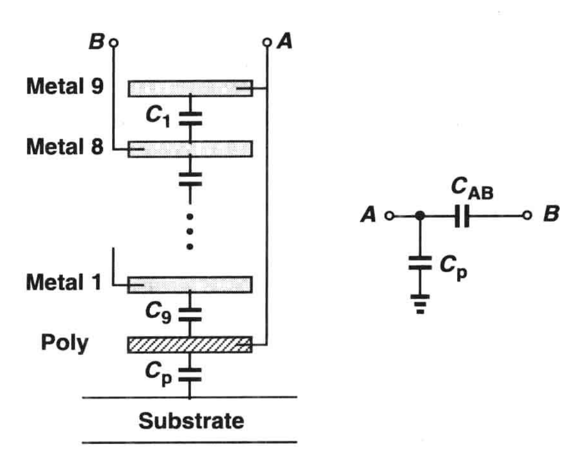


Figure 7.79 *Bottom-plate parasitic capacitance.*

Example 7.37

We wish to employ capacitive coupling at the input of a stage that has an input capacitance of C_{in} (Fig. 7.80). Determine the additional input capacitance resulting from the coupling capacitor. Assume $C_p = 0.1C_c$.

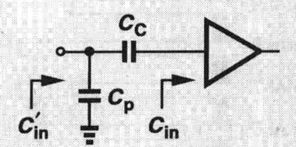


Figure 7.80 *Choice of input coupling capacitance value.*

Solution:

To minimize signal attenuation, C_c must be much greater than C_{in}, e.g., $C_c \approx 5C_{in}$. Thus, $C_p = 0.5C_{in}$, yielding

$$C'_{in} = \frac{C_c C_{in}}{C_c + C_{in}} + 0.5C_{in} \tag{7.123}$$

$$= \frac{4}{3}C_{in}. \tag{7.124}$$

That is, the input capacitance is raised by more than 30%.

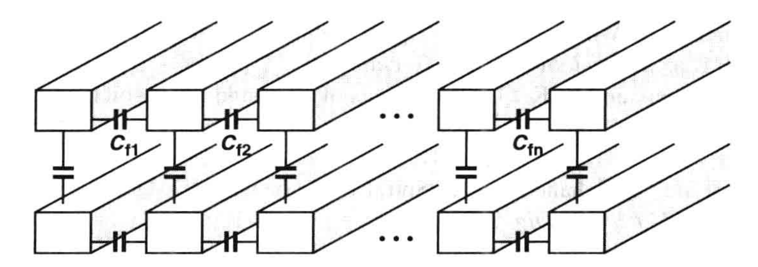


Figure 7.81 *Fringe capacitor structure.*

To alleviate the above issue, only a few top metal layers can be utilized. For example, a structure consisting of metal 9 through metal 4 has a density of 660 aF/μm^2 and a parasitic of 18 aF/μm^2, i.e., 2.7%. Of course, the lower density translates to a larger area and more complex routing of signals.

An alternative geometry utilizes the lateral electric field between adjacent metal lines to achieve a high capacitance density. Illustrated in Fig. 7.81, this "fringe" capacitor consists of narrow metal lines with the minimum allowable spacing. This structure is described in Chapter 8.

REFERENCES

[1] J. Craninckx and M. S. J. Steyaert, "A 1.8 GHz CMOS Low Phase Noise Voltage-Controlled Oscillator with Prescaler," *IEEE Journal of Solid-State Circuits,* vol. 30, pp. 1474–1482, Dec. 1995.

[2] S. Jenei, B. K. J. C. Nauwelaers, and S. Decoutere, "Physics-Based Closed-Form Inductance Expressions for Compact Modeling of Integrated Spiral Inductors," *IEEE Journal of Solid-State Circuits,* vol. 37, pp. 77–80, Jan. 2002.

[3] S. S. Mohan et al., "Simple Accurate Expressions for Planar Spiral Inductances," *IEEE J. Solid-State Circuits,* vol. 34, pp. 1419–1424, Oct. 1999.

[4] A. Niknejad and R. G. Meyer, "Analysis, Design, and Optimization of Spiral Inductors and Transformers for Si RF ICs," *IEEE J. Solid-State Circuits,* vol. 33, pp. 1470–1481, Oct. 1998.

[5] M. Kraemer, D. Dragomirescu, and R. Plana, "Accurate Electromagnetic Simulation and Measurement of Millimeter-Wave Inductors in Bulk CMOS Technology," *IEEE Topical Meeting on Silison Monolithic Integrated Circuits in RF Systems,* pp. 61–64, Jan. 2010.

[6] A. Zolfaghari, A. Y. Chan, and B. Razavi, "Stacked Inductors and 1-to-2 Transformers in CMOS Technology," *IEEE Journal of Solid-State Circuits,* vol. 36, pp. 620–628, April 2001.

[7] W. B. Kuhn, "Approximate Analytical Modeling of Current Crowding Effects in Multi-Turn Spiral Inductors," *IEEE RFIC Symp. Dig. Tech. Papers,* pp. 271–274, June 2000.

[8] W. B. Kuhn and N. M. Ibrahim, "Analysis of Current Crowding Effects in Multiturn Spiral Inductors," *IEEE Tran. MTT,* vol. 49, pp. 31–39, Jan. 2001.

[9] C. S. Yen, Z. Fazarinc, and R. Wheeler, "Time-Domain Skin-Effect Model for Transient Analysis of Lossy Transmission Lines," *Proc. of IEEE,* vol. 70, pp. 750–757, July 1982.

[10] Y. Cao et al., "Frequency-Independent Equivalent Circuit Model of On-Chip Spiral Inductors," *Proc. CICC,* pp. 217–220, May 2002.

[11] M. Danesh et al., "A Q-Factor Enhancement Technique for MMIC Inductors," *Proc. IEEE Radio Frequency Integrated Circuits Symp.,* pp. 217–220, April 1998.

[12] N. M. Neihart et al., "Twisted Inductors for Low Coupling Mixed-signal and RF Applications," *Proc. CICC,* pp. 575–578, Sept. 2008.

[13] C. P. Yue and S. S. Wong, "On-Chip Spiral Inductors with Patterned Ground Shields for Si-Based RF ICs," *IEEE J. Solid-State Circuits*, vol. 33, pp. 743–751, May 1998.
[14] S.-M. Yim, T. Chen, and K. K. O, "The Effects of a Ground Shield on the Characteristics and Performance of Spiral Inductors," *IEEE J. Solid-State Circuits*, vol. 37, pp. 237–245, Feb. 2002.
[15] Y. E. Chen et al., "Q-Enhancement of Spiral Inductor with N^+-Diffusion Patterned Ground Shields," *IEEE MTT Symp. Dig. Tech. Papers*, pp. 1289–1292, 2001.
[16] J. R. Long, "Monolithic Transformers for Silicon RF IC Design," *IEEE J. Solid-State Circuits*, vol. 35, pp. 1368–1383, Sept. 2000.
[17] J. R. Long and M. A. Copeland, "The Modeling, Characterization, and Design of Monolithic Inductors for Silicon RF ICs," *IEEE J. Solid-State Circuits*, vol. 32, pp. 357–369, March 1997.
[18] S. Ramo, J. R. Whinnery, and T. Van Duzer, *Fields and Waves in Communication Electronics*, Third Edition, New York: Wiley, 1994.

PROBLEMS

7.1. Extend Eq. (7.1) to an N-turn spiral and show that L_{tot} contains $N(N+1)/2$ terms.

7.2. Prove that the Q of the circuit shown in Fig. 7.32(a) is given by Eq. (7.62).

7.3. Prove that for an N-turn spiral inductor, the equivalent interwinding capacitance is given by

$$C_{eq} = \frac{C_1 + \cdots + C_{N^2-1}}{(N^2-1)^2}. \tag{7.125}$$

7.4. Using Eq. (7.62), compare the Q's of the circuits shown in Figs. 7.37(b) and (d).

7.5. Consider the magnetic fields produced by the inductors in Fig. 7.41. Which topology creates less net magnetic field at a point far from the circuit but on its line of symmetry?

7.6. Repeat Example 7.13 for a 5-nH inductor using a linewidth of 5 μm, a line spacing of 0.5 μm, and four turns. Do the results depend much on the outer diameter?

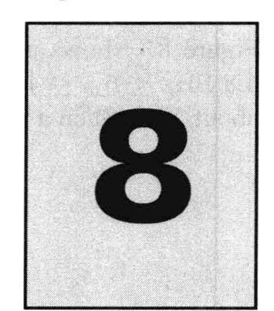
7.7. For the circuit of Fig. 7.28(a), compute Y_{11} and find the parallel equivalent resistance. Is the result the same as that shown in Eq. (7.55)?

7.8. Repeat Example 7.19 for four turns. Is it possible to find an expression for N turns?

7.9. Find the input impedance, Z_{in}, in Fig. 7.50.

7.10. Using the capacitance data in Table 7.1, repeat Example 7.25 for an inductor realized as a stack of four metal layers. Assume the inductance is about 3.5 times that of one spiral.

CHAPTER



OSCILLATORS

In our study of RF transceivers in Chapter 4, we noted the extensive use of oscillators in both the transmit and receive paths. Interestingly, in most systems, one input of every mixer is driven by a periodic signal, hence the need for oscillators. This chapter deals with the analysis and design of oscillators. The outline is shown below.

General Principles	Voltage-Controlled Oscillators	Phase Noise	Quadrature VCOs
Feedback View	Tuning Limitations	Effect of Phase Noise	Coupling into an Oscillator
One-Port View	Effect of Varactor Q	Analysis Approach I	Basic Topology
Cross-Coupled Oscillator	VCOs with Wide Tuning Range	Analysis Approach II	Properties of Quadrature Oscillators
Three-Point Oscillators		Noise of Bias Current	Improved Topologies
		VCO Design Procedure	
		Low-Noise VCOs	

8.1 PERFORMANCE PARAMETERS

An oscillator used in an RF transceiver must satisfy two sets of requirements: (1) system specifications, e.g., the frequency of operation and the "purity" of the output, and (2) "interface" specifications, e.g., drive capability or output swing. In this section, we study the oscillator performance parameters and their role in the overall system.

Frequency Range An RF oscillator must be designed such that its frequency can be varied (tuned) across a certain range. This range includes two components: (1) the system specification; for example, a 900-MHz GSM direct-conversion receiver may tune the LO from 935 MHz to 960 MHz; (2) additional margin to cover process and temperature variations and errors due to modeling inaccuracies. The latter component typically amounts to several percent.

Example 8.1

A direct-conversion transceiver is designed for the 2.4-GHz and 5-GHz wireless bands. If a single LO must cover both bands, what is the minimum acceptable tuning range?

Solution:

Figure 8.1 shows a possible arrangement for covering both bands. For the lower band, 4.8 GHz $\leq f_{LO} \leq$ 4.96 GHz. Thus, we require a total tuning range of 4.8 GHz to 5.8 GHz, about 20%. Such a wide tuning range is relatively difficult to achieve in LC oscillators.

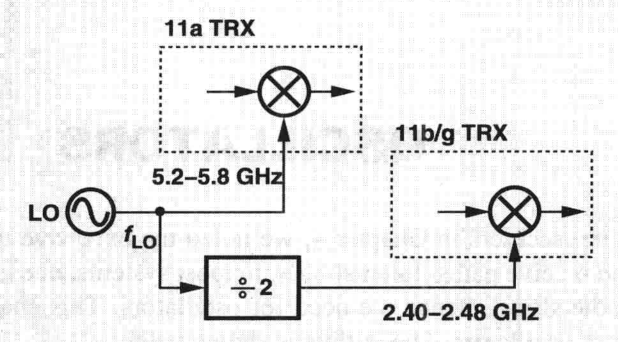


Figure 8.1 *LO path of a dual-band transceiver.*

The actual frequency range of an oscillator may also depend on whether quadrature outputs are required and/or injection pulling is of concern (Chapter 4). A direct-conversion transceiver employs quadrature phases of the carrier, necessitating that either the oscillator directly generate quadrature outputs or it run at twice the required frequency so that a ÷2 stage can produce such outputs. For example, in the hybrid topology of Fig. 8.1, the LO must still provide quadrature phases in the 5-GHz range—but it is prone to injection pulling by the PA output. We address the problem of quadrature generation in Section 8.11.

How high a frequency can one expect of a CMOS oscillator? While oscillation frequencies as high as 300 GHz have been demonstrated [1], in practice, a number of serious trade-offs emerge that become much more pronounced at higher operation frequencies. We analyze these trade-offs later in this chapter.

Output Voltage Swing As exemplified by the arrangement shown in Fig. 8.1, the oscillators in an RF system drive mixers and frequency dividers. As such, they must produce sufficiently large output swings to ensure nearly complete switching of the transistors in the subsequent stages. Furthermore, as studied in Section 8.7, excessively low output swings exacerbate the effect of the internal noise of the oscillator. With a 1-V supply, a typical single-ended swing may be around 0.6 to 0.8 V_{pp}. A buffer may follow the oscillator to amplify the swings and/or drive the subsequent stage.

Drive Capability Oscillators may need to drive a large load capacitance. Figure 8.2 depicts a typical arrangement for the receive path. In addition to the downconversion mixers, the oscillator must also drive a frequency divider, denoted by a $\div N$ block. This is because a loop called the "frequency synthesizer" must precisely control the frequency of

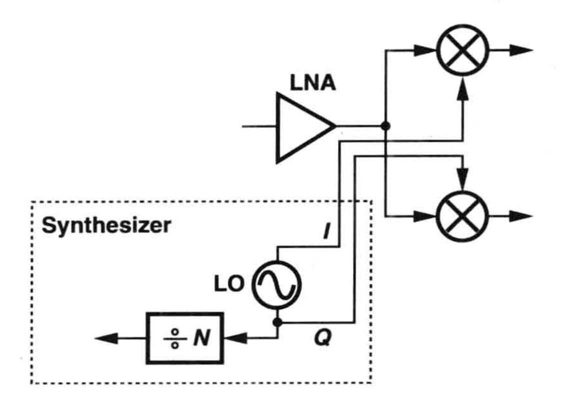


Figure 8.2 *Circuits loading the LO.*

the oscillator, requiring a divider (Chapter 10). In other words, the LO must drive the input capacitance of at least one mixer and one divider.

Interestingly, typical mixers and dividers exhibit a trade-off between the minimum LO swing with which they can operate properly and the capacitance that they present at their LO port. This can be seen in the representative stage shown in Fig. 8.3(a), wherein it is desirable to switch M_1 and M_2 as abruptly as possible (Chapter 6). To this end, we can select large LO swings so that $V_{GS1} - V_{GS2}$ rapidly reaches a large value, turning off one transistor [Fig. 8.3(b)]. Alternatively, we can employ smaller LO swings but *wider* transistors so that they steer their current with a smaller differential input.

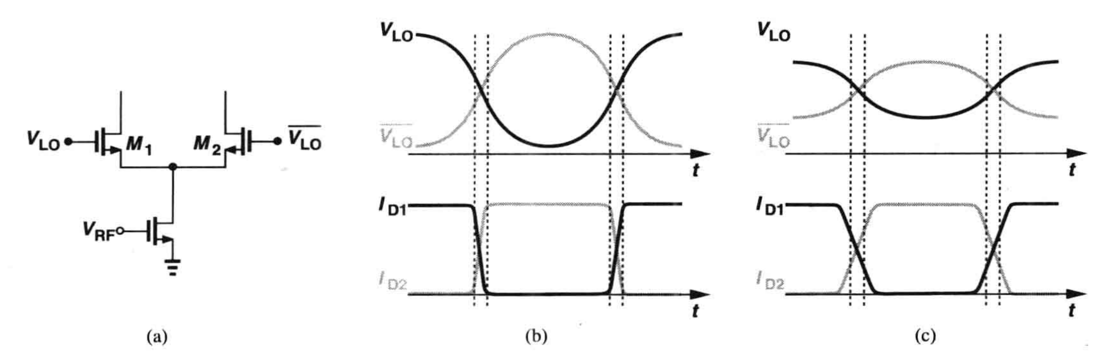


Figure 8.3 *(a) Representative LO path of mixers and dividers, (b) current-steering with large LO swings, (c) current-steering with small LO swings but wide transistors.*

The issue of capacitive loading becomes more serious in transmitters. As explained in Example 6.37, the PA input capacitance "propagates" to the LO port of the upconversion mixers.

To alleviate the loading presented by mixers and dividers and perhaps amplify the swings, we can follow the LO with a buffer, e.g., a differential pair. Note that in Fig. 8.2, *two* buffers are necessary for the quadrature phases. The buffers consume additional power and may require inductive loads—owing to speed limitations or the need for swings above the supply voltage (Chapter 6). The additional inductors complicate the layout and the routing of the high-frequency signals.

Example 8.2

Prove that the LO port of downconversion mixers presents a mostly capacitive impedance, whereas that of upconversion mixers also contains a resistive component.

Solution:

Consider the simplified model shown in Fig. 8.4. Here, R_p represents a physical load resistor in a downconversion mixer, forming a low-pass filter with C_L. In an upconversion mixer, on the other hand, R_p models the equivalent parallel resistance of a load inductor at resonance. In Problem 8.1, we will compute the input admittance of the circuit and show that the real part reduces to the following form:

$$Re\{Y_{in}\} = \frac{[(1 + g_m R_p)C_{GD} + g_m R_p C_L] R_p C_{GD} \omega^2}{1 + R_p^2 (C_{GD} + C_L)^2 \omega^2}. \tag{8.1}$$

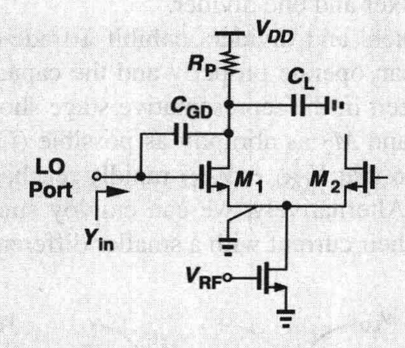


Figure 8.4 *Input admittance of differential pair.*

In a downconversion mixer, the −3-dB bandwidth at the output node is commensurate with the *channel* bandwidth and hence very small. That is, we can assume $R_p C_L$ is very large, simplifying Eq. (8.1) to

$$Re\{Y_{in}\} \approx \frac{g_m C_{GD} C_L}{(C_{GD} + C_L)^2}. \tag{8.2}$$

It is also reasonable to assume that $C_L \gg C_{GD}$ in a downconversion mixer, arriving at the following *parallel* resistive component seen at the input:

$$R_{in} \approx \frac{1}{g_m} \frac{C_L}{C_{GD}}. \tag{8.3}$$

Note that this is the resistance seen in parallel with the input, but only for a fraction of the period (when M_1 and M_2 are around equilibrium). For example, if $1/g_m \approx 100\ \Omega$, $C_L = 1$ pF, and $C_{GD} = 5$ fF, then $R_{in} = 20$ kΩ, a relatively large value.

In an upconversion mixer, Eq. (8.1) may yield a substantially lower input resistance. For example, if $g_m R_p = 2$, $R_p = 200\ \Omega$, $C_{GD} = 5$ fF, $C_L = 20$ fF, and $\omega = 2\pi \times (10\,\text{GHz})$, then $R_{in} = 5.06$ kΩ. (In practice, C_L is nulled by the load inductor.) This resistive component loads the LO, degrading its performance.

Phase Noise The spectrum of an oscillator in practice deviates from an impulse and is "broadened" by the noise of its constituent devices. Called "phase noise," this phenomenon has a profound effect on RF receivers and transmitters (Section 8.7). Unfortunately, phase noise bears direct trade-offs with the tuning range and power dissipation of oscillators, making the design more challenging. Since the phase noise of LC oscillators is inversely proportional to the Q of their tank(s), we will pay particular attention to factors that degrade the Q.

Output Waveform What is the desired output waveform of an RF oscillator? Recall from the analysis of mixers in Chapter 6 that *abrupt* LO transitions reduce the noise and increase the conversion gain. Moreover, effects such as direct feedthrough are suppressed if the LO signal has a 50% duty cycle. Sharp transitions also improve the performance of frequency dividers (Chapter 10). Thus, the ideal LO waveform in most cases is a square wave.

In practice, it is difficult to generate square LO waveforms. This is primarily because the LO circuitry itself and the buffer(s) following it typically incorporate (narrowband) resonant loads, thereby attenuating the harmonics. For this reason, as illustrated in Fig. 8.3, the LO amplitude is chosen large and/or the switching transistors wide so as to approximate abrupt current switching.

A number of considerations call for *differential* LO waveforms. First, as observed in Chapter 6, *balanced* mixers outperform unbalanced topologies in terms of gain, noise, and dc offsets. Second, the leakage of the LO to the input is generally smaller with differential waveforms.

Supply Sensitivity The frequency of an oscillator may vary with the supply voltage, an undesirable effect because it translates supply noise to frequency (and phase) noise. For example, external or internal voltage regulators may suffer from substantial flicker noise, which cannot be easily removed by bypass capacitors due to its low-frequency contents. This noise therefore modulates the oscillation frequency (Fig. 8.5).

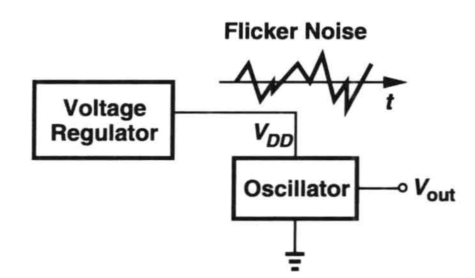


Figure 8.5 *Effect of regulator noise on an oscillator.*

Power Dissipation The power drained by the LO and its buffer(s) proves critical in some applications as it trades with the phase noise and tuning range. Thus, many techniques have been introduced that lower the phase noise for a given power dissipation.

8.2 BASIC PRINCIPLES

An oscillator generates a periodic output. As such, the circuit must involve a self-sustaining mechanism that allows its own noise to grow and eventually become a periodic signal.

8.2.1 Feedback View of Oscillators

An oscillator may be viewed as a "badly-designed" negative-feedback amplifier—so badly-designed that it has a zero or negative phase margin. While the art of oscillator design entails much more than an unstable amplifier, this view provides a good starting point for our study. Consider the simple linear negative-feedback system depicted in Fig. 8.6, where

$$\frac{Y}{X}(s) = \frac{H(s)}{1 + H(s)}. \tag{8.4}$$

What happens if at a sinusoidal frequency, ω_1, $H(s = j\omega_1)$ becomes equal to -1? The gain from the input to the output goes to infinity, allowing the circuit to amplify a noise component at ω_1 indefinitely. That is, the circuit can sustain an output at ω_1. From another point of view, the closed-loop system exhibits two imaginary poles given by $\pm j\omega_1$.

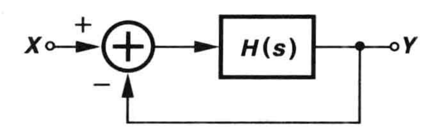


Figure 8.6 *Negative feedback system.*

Example 8.3

For the above system to oscillate, must the noise at ω_1 appear at the input?

Solution:

No, the noise can be anywhere in the loop. For example, consider the system shown in Fig. 8.7, where the noise N appears in the feedback path. Here,

$$Y(s) = \frac{H_1(s)}{1 + H_1(s)H_2(s)H_3(s)}X(s) + \frac{H_1(s)H_3(s)}{1 + H_1(s)H_2(s)H_3(s)}N(s). \tag{8.5}$$

Thus, if the loop transmission, $H_1H_2H_3$, approaches -1 at ω_1, N is also amplified indefinitely.

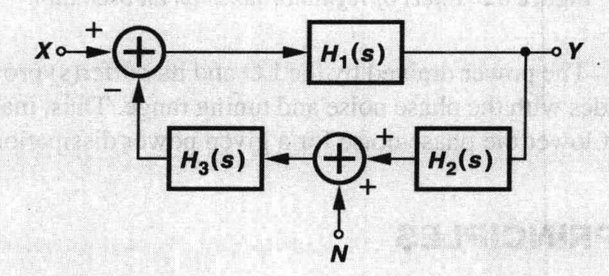


Figure 8.7 *Negative feedback systems with two injection points.*

The above example leads to a general and powerful analytical point: in the small-signal model of an oscillator, the impedance seen between *any* two nodes (one of which is not ground) in the signal path goes to infinity at the oscillation frequency, ω_1, because a noise current at ω_1 injected between these two nodes produces an infinitely large swing. This observation can be used to determine the oscillation condition and frequency.

Example 8.4

Derive an expression for Y/X in Fig. 8.6 in the vicinity of $\omega = \omega_1$ if $H(j\omega_1) = -1$.

Solution:

If $\omega = \omega_1 + \Delta\omega$, we can approximate $H(j\omega)$ by the first two terms in its Taylor series:

$$H[j(\omega_1 + \Delta\omega)] \approx H(j\omega_1) + \Delta\omega \frac{dH(j\omega_1)}{d\omega}. \tag{8.6}$$

Since $H(j\omega_1) = -1$, we have

$$\frac{Y}{X}[j(\omega_1 + \Delta\omega)] = \frac{H(j\omega_1) + \Delta\omega \dfrac{dH(j\omega_1)}{d\omega}}{\Delta\omega \dfrac{dH(j\omega_1)}{d\omega}} \tag{8.7}$$

$$\approx \frac{H(j\omega_1)}{\Delta\omega \dfrac{dH(j\omega_1)}{d\omega}} \tag{8.8}$$

$$\approx \frac{-1}{\Delta\omega \dfrac{dH(j\omega_1)}{d\omega}}. \tag{8.9}$$

As expected, $Y/X \to \infty$ as $\Delta\omega \to 0$, with a "sharpness" proportional to $dH/d\omega$.

Since $H(s)$ is a complex function, the condition $H(j\omega_1) = -1$ can equivalently be expressed as

$$|H(s = j\omega_1)| = 1 \tag{8.10}$$

$$\angle H(s = j\omega_1) = 180^\circ, \tag{8.11}$$

which are called "Barkhausen's criteria" for oscillation. Let us examine these two conditions to develop more insight. We recognize that a signal at ω_1 experiences a gain of unity and a phase shift of 180° as it travels through $H(s)$ [Fig. 8.8(a)]. Bearing in mind that the system is originally designed to have *negative* feedback (as denoted by the input subtractor), we conclude that the signal at ω_1 experiences a *total* phase shift of 360° [Fig. 8.8(b)] as it travels around the loop. This is, of course, to be expected: for the circuit to reach steady state, the signal returning to A must exactly coincide with the signal that started at A.

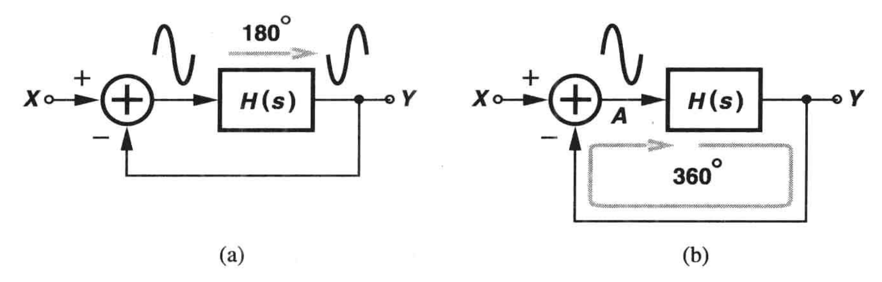


Figure 8.8 *Barkhausen's phase shift criterion viewed as (a)* 180° *frequency-dependent phase shift due to* $H(s)$, *(b)* 360° *total phase shift.*

We call $\angle H(j\omega_1)$ a "frequency-dependent" phase shift to distinguish it from the 180° phase due to negative feedback.

The above point can also be viewed as follows. Even though the system was originally configured to have negative feedback, $H(s)$ is so "sluggish" that it contributes an additional phase shift of 180° at ω_1, thereby creating *positive* feedback at this frequency.

What is the significance of $|H(j\omega_1)| = 1$? For a noise component at ω_1 to "build up" as it circulates around the loop with positive feedback, the loop gain must be at least unity. Figure 8.9 illustrates the "startup" of the oscillator if $|H(j\omega_1)| = 1$ and $\angle H(j\omega_1) = 180^\circ$. An input at ω_1 propagates through $H(s)$, emerging unattenuated but inverted. The result is *subtracted* from the input, yielding a waveform with twice the amplitude. This growth continues with time. We call $|H(j\omega_1)| = 1$ the "startup" condition.

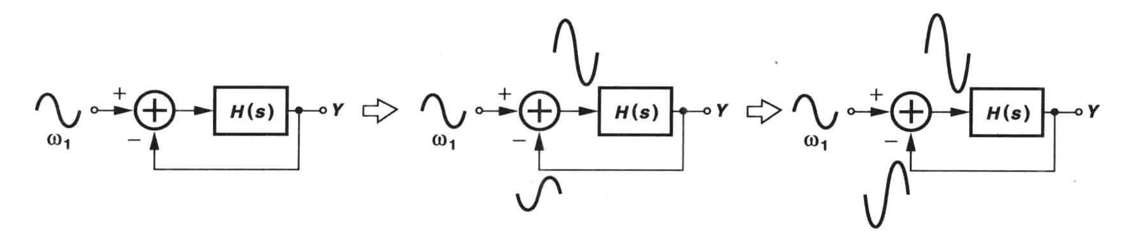


Figure 8.9 *Successive snapshots of an oscillator during startup.*

What happens if $|H(j\omega_1)| > 1$ and $\angle H(j\omega_1) = 180^\circ$? The growth shown in Fig. 8.9 still occurs but at a faster rate because the returning waveform is amplified by the loop. Note that the closed-loop poles now lie in the right half plane. The amplitude growth eventually ceases due to circuit nonlinearities. We elaborate on these points later in this chapter.

Example 8.5

Can a two-pole system oscillate?

Solution:

Suppose the system exhibits two coincident real poles at ω_p. Figure 8.10(a) shows an example, where two cascaded common-source stages constitute $H(s)$ and $\omega_p = (R_1C_1)^{-1}$.

Example 8.5 *(Continued)*

This system cannot satisfy both of Barkhausen's criteria because the phase shift associated with each stage reaches 90° only at $\omega = \infty$, but $|H(\infty)| = 0$. Figure 8.10(b) plots $|H|$ and $\angle H$ as a function of frequency, revealing no frequency at which both conditions are met. Thus, the circuit cannot oscillate.

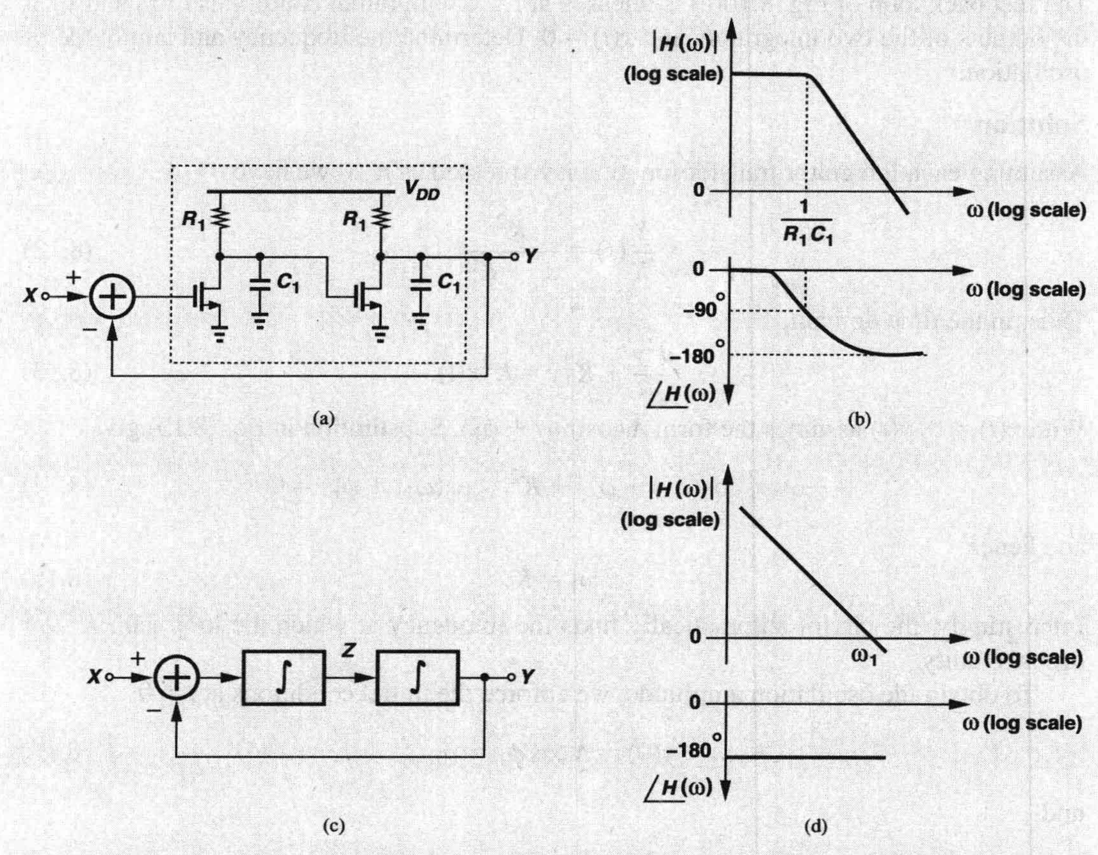


Figure 8.10 *(a) Two CS stages in feedback, (b) loop transmission response, (c) two integrators in feedback, (d) loop transmission response.*

But, what if both poles are located at the *origin*? Realized as two ideal integrators in a loop [Fig. 8.10(c)], such a circuit does oscillate because each integrator contributes a phase shift of −90° at any nonzero frequency. Shown in Fig. 8.10(d) are $|H|$ and $\angle H$ for this system.

Our study thus far allows us to predict the frequency of oscillation: we seek the frequency at which the total phase shift around the loop is 360° and determine whether the loop gain reaches unity at this frequency. (An exception is described in the example below.) This calculation, however, does not predict the oscillation *amplitude*. In a perfectly

linear loop, the oscillation amplitude is simply given by the initial conditions residing on the storage elements if the loop gain is equal to unity at the oscillation frequency. The following example illustrates this point.

Example 8.6

The feedback loop of Fig. 8.10(c) is released at $t = 0$ with initial conditions of z_0 and y_0 at the outputs of the two integrators and $x(t) = 0$. Determine the frequency and amplitude of oscillation.

Solution:

Assuming each integrator transfer function is expressed as K/s, we have

$$\frac{Y}{X}(s) = \frac{K^2}{s^2 + K^2}. \tag{8.12}$$

Thus, in the time domain,

$$\frac{d^2y}{dt^2} + K^2 y = K^2 x(t). \tag{8.13}$$

With $x(t) = 0$, $y(t)$ assumes the form $A\cos(\omega_1 t + \phi_1)$. Substitution in Eq. (8.13) gives

$$-A\omega_1^2 \cos(\omega_1 t + \phi_1) + K^2 A\cos(\omega_1 t + \phi_1) = 0 \tag{8.14}$$

and hence

$$\omega_1 = K. \tag{8.15}$$

Interestingly, the circuit automatically finds the frequency at which the loop gain K^2/ω^2 drops to unity.

To obtain the oscillation amplitude, we enforce the initial conditions at $t = 0$:

$$y(0) = A\cos\phi_1 = y_0, \tag{8.16}$$

and

$$z(0) = \frac{1}{K}\frac{dy}{dt}|_{t=0} \tag{8.17}$$

$$= -A\sin\phi_1 = z_0. \tag{8.18}$$

It follows from Eqs. (8.16) and (8.18) that

$$\tan\phi_0 = -\frac{z_0}{y_0} \tag{8.19}$$

$$A = \sqrt{z_0^2 + y_0^2}. \tag{8.20}$$

Example 8.6 *(Continued)*

Why does the circuit not oscillate at frequencies *below* $\omega_1 = K$? It appears that the loop has enough gain and a phase shift of 180° at these frequencies. As mentioned earlier, oscillation build-up occurs with a loop gain of greater than unity only if the closed-loop system contains poles in the right-half plane. This is not the case for the two-integrator loop: Y/X can have only poles on the imaginary axis, failing to produce oscillation if $s = j\omega \neq jK$.

Other oscillators behave differently from the two-integrator loop: they may begin to oscillate at a frequency at which the loop gain is *higher* than unity, thereby experiencing an exponential growth in their output amplitude. In an actual circuit, the growth eventually stops due to the saturating behavior of the amplifier(s) in the loop. For example, consider the cascade of three CMOS inverters depicted in Fig. 8.11 (called a "ring oscillator"). If the circuit is released with X, Y, and Z at the trip point of the inverters, then each stage operates as an amplifier, leading to an oscillation frequency at which each inverter contributes a frequency-dependent phase shift of 60°. (The three inversions make this a negative-feedback loop at low frequencies.) With the high loop gain, the oscillation amplitude grows exponentially until the transistors enter the triode region at the peaks, thus lowering the gain. In the steady state, the output of each inverter swings from nearly zero to nearly V_{DD}.

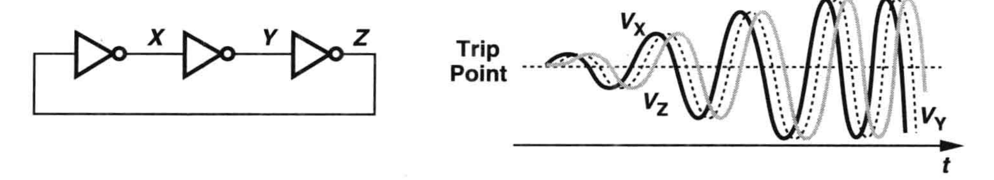


Figure 8.11 *Ring oscillator and its waveforms.*

In most oscillator topologies of interest to us, the voltage swings are defined by the saturating behavior of differential pairs. The following example elaborates on this point.

Example 8.7

The inductively-loaded differential pair shown in Fig. 8.12(a) is driven by a large input sinusoid at $\omega_0 = 1/\sqrt{L_1 C_1}$. Plot the output waveforms and determine the output swing.

(Continues)

Example 8.7 *(Continued)*

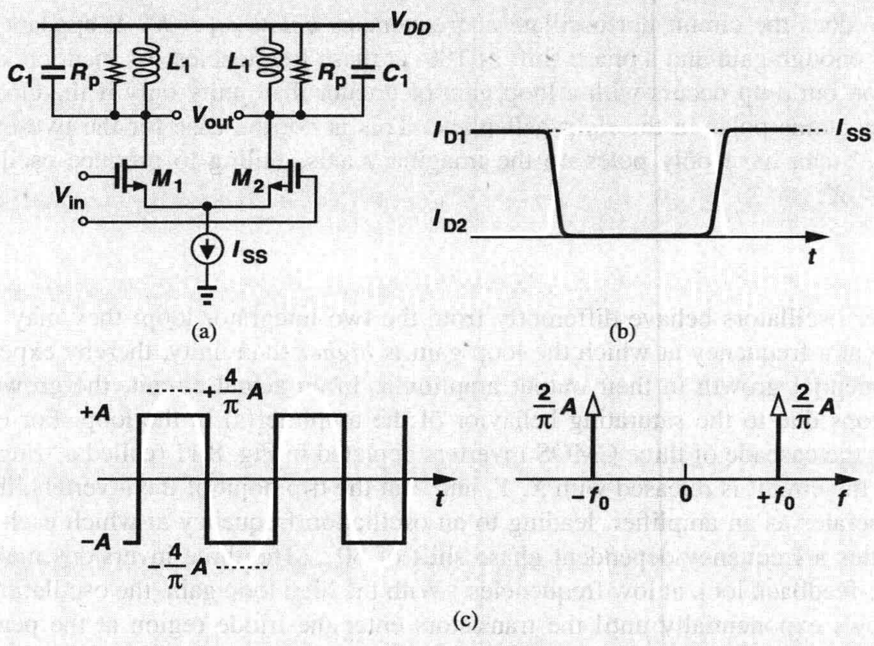


Figure 8.12 *(a) Differential pair with tank loads, (b) drain currents, (c) first harmonic of a square wave.*

Solution:

With large input swings, M_1 and M_2 experience complete switching in a short transition time, injecting nearly square current waveforms into the tanks [Fig. 8.12(b)]. Each drain current waveform has an average of $I_{SS}/2$ and a peak amplitude of $I_{SS}/2$. The first harmonic of the current is multiplied by R_p, whereas higher harmonics are attenuated by the tank selectivity. Recall from the Fourier expansion of a square wave of peak amplitude A (with 50% duty cycle) that the first harmonic exhibits a peak amplitude of $(4/\pi)A$ (slightly *greater* than A) [Fig. 8.12(c)]. The peak single-ended output swing is therefore equal to $(4/\pi)(I_{SS}/2)R_p = 2I_{SS}R_p/\pi$, yielding a peak differential output swing of

$$V_{out} = \frac{4}{\pi}I_{SS}R_p. \tag{8.21}$$

If interested in carrying out this calculation in the frequency domain, the reader is cautioned that the spectrum of the first harmonic contains two impulses, each having an area of $(2/\pi)A$ [not $(4/\pi)A$] [Fig. 8.12(c)].

8.2.2 One-Port View of Oscillators

In the previous section, we considered oscillators as negative-feedback systems that experience sufficient positive feedback at some frequency. An alternative perspective views

oscillators as two one-port components, namely, a lossy resonator and an active circuit that cancels the loss. This perspective provides additional insight and is described in this section.

Suppose, as shown in Fig. 8.13(a), a current impulse, $I_0\delta(t)$, is applied to a lossless tank. The impulse is entirely absorbed by C_1 (why?), generating a voltage of I_0/C_1. The charge on C_1 then begins to flow through L_1, and the output voltage falls. When V_{out} reaches zero, C_1 carries no energy but L_1 has a current equal to $L_1 dV_{out}/dt$, which charges C_1 in the opposite direction, driving V_{out} toward its negative peak. This periodic exchange of energy between C_1 and L_1 continues indefinitely, with an amplitude given by the strength of the initial impulse.

Now, let us assume a lossy tank. Depicted in Fig. 8.13(b), such a circuit behaves similarly except that R_p drains and "burns" some of the capacitor energy in every cycle, causing an exponential decay in the amplitude. We therefore surmise that, if an active circuit replenishes the energy lost in each period, then the oscillation can be sustained. In fact, we predict that an active circuit exhibiting an input resistance of $-R_p$ can be attached across the tank to cancel the effect of R_p, thereby recreating the ideal scenario shown in Fig. 8.13(a). Illustrated in Fig. 8.13(c), the resulting topology is called a "one-port oscillator."

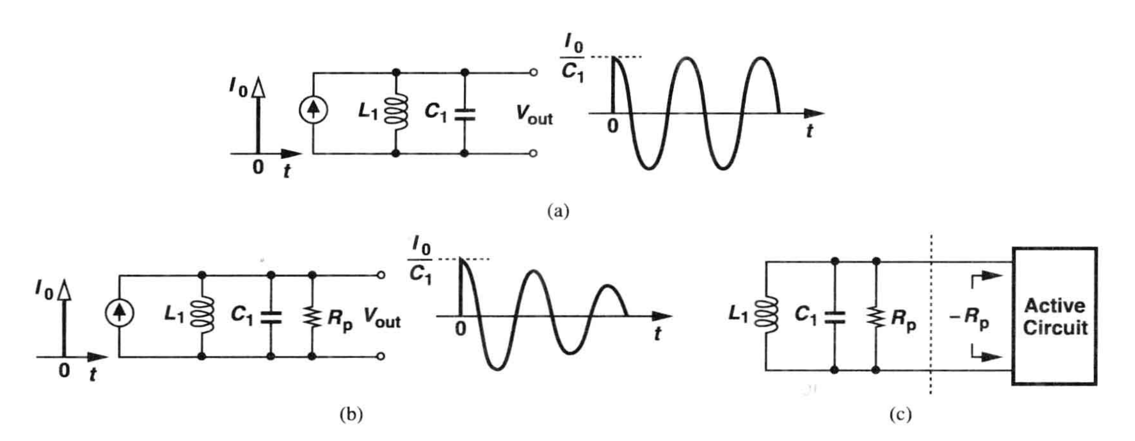


Figure 8.13 *(a) Response of an ideal tank to an impulse, (b) response of a lossy tank to an impulse, (c) cancellation of loss by negative resistance.*

Example 8.8

A student who remembers that loss in a tank results in noise postulates that, if the circuit of Fig. 8.13(c) resembles the ideal lossless topology, then it must also exhibit zero noise. Is that true?

Solution:

No, it is not. Resistance R_p and the active circuit still generate their own (uncorrelated) noise. We return to this point in Section 8.7.

How can a circuit present a negative (small-signal) input resistance? Figure 8.14(a) shows an example, where two capacitors are tied from the gate and drain of a transistor to its

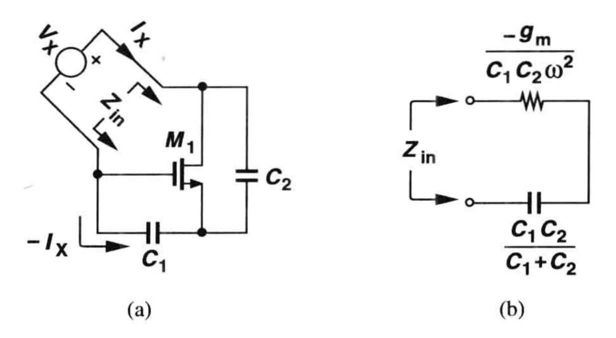


Figure 8.14 *(a) Circuit providing negative resistance, (b) equivalent circuit.*

source. The impedance Z_{in} can be obtained by noting that C_1 carries a current equal to $-I_X$, generating a gate-source voltage of $-I_X/(C_1 s)$ and hence a drain current of $-I_X g_m/(C_1 s)$. The difference between I_X and the drain current flows through C_2, producing a voltage equal to $[I_X + I_X g_m/(C_1 s)]/(C_2 s)$. This voltage must be equal to $V_{GS} + V_X$:

$$-\frac{I_X}{C_1 s} + V_X = \left(I_X + I_X \frac{g_m}{C_1 s}\right) \frac{1}{C_2 s}. \tag{8.22}$$

It follows that

$$\frac{V_X}{I_X}(s) = \frac{1}{C_1 s} + \frac{1}{C_2 s} + \frac{g_m}{C_1 C_2 s^2}. \tag{8.23}$$

For a sinusoidal input, $s = j\omega$,

$$\frac{V_X}{I_X}(j\omega) = \frac{1}{jC_1\omega} + \frac{1}{jC_2\omega} - \frac{g_m}{C_1 C_2 \omega^2}. \tag{8.24}$$

Thus, the input impedance can be viewed as a series combination of C_1, C_2, and a *negative* resistance equal to $-g_m/(C_1 C_2 \omega^2)$ [Fig. 8.14(b)]. Interestingly, the negative resistance *varies* with frequency.

Having found a negative resistance, we can now attach it to a lossy tank so as to construct an oscillator. Since the capacitive component in Eq. (8.24) can become part of the tank, we simply connect an inductor to the negative-resistance port (Fig. 8.15), seeking the condition for oscillation. In this case, it is simpler to model the loss of the inductor by a series resistance, R_S. The circuit oscillates if

$$R_S = \frac{g_m}{C_1 C_2 \omega^2}. \tag{8.25}$$

Under this condition, the circuit reduces to L_1 and the series combination of C_1 and C_2, exhibiting an oscillation frequency of

$$\omega_{osc} = \frac{1}{\sqrt{L_1 \dfrac{C_1 C_2}{C_1 + C_2}}}. \tag{8.26}$$

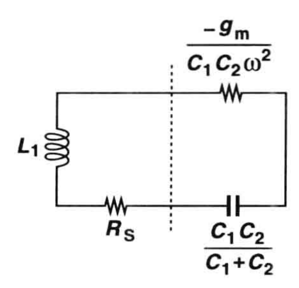


Figure 8.15 *Connection of lossy inductor to negative-resistance circuit.*

Example 8.9

Express the oscillation condition in terms of inductor's parallel equivalent resistance, R_p, rather than R_S.

Solution:

Recall from Chapter 2 that, if $Q > 3$, the series combination can be transformed to a parallel combination and

$$\frac{L_1\omega}{R_S} \approx \frac{R_p}{L_1\omega}. \tag{8.27}$$

Thus,

$$\frac{L_1^2\omega^2}{R_p} = \frac{g_m}{C_1C_2\omega^2}. \tag{8.28}$$

Moreover, we can replace ω^2 with the value given by Eq. (8.26), arriving at the startup condition:

$$g_mR_p = \frac{(C_1 + C_2)^2}{C_1C_2} \tag{8.29}$$

$$= \frac{C_1}{C_2} + \frac{C_2}{C_1} + 2. \tag{8.30}$$

As expected, for oscillation to occur, the transistor in Fig. 8.14(a) must provide sufficient "strength" (transconductance). In fact, (8.30) implies that the minimum allowable g_m is obtained if $C_1 = C_2$. That is, $g_mR_p \geq 4$.

8.3 CROSS-COUPLED OSCILLATOR

In this section, we develop an LC oscillator topology that, owing to its robust operation, has become the dominant choice in RF applications. We begin the development with a feedback system, but will discover that the result also lends itself to the one-port view described in Section 8.2.2.

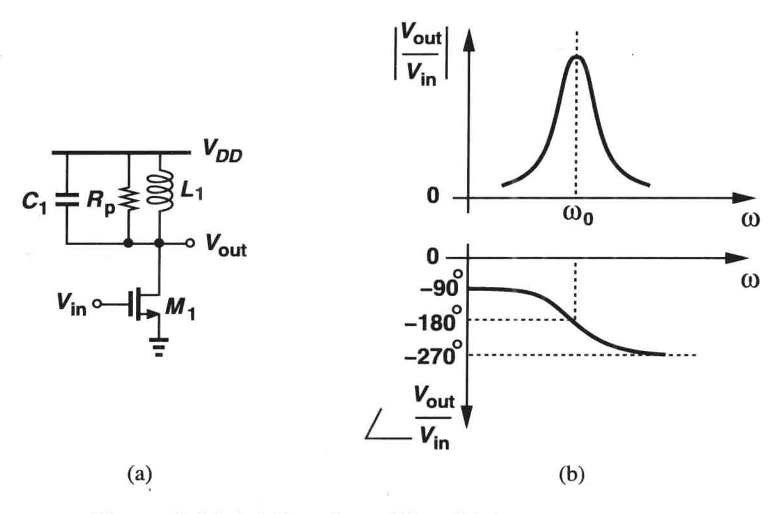


Figure 8.16 *(a) Tuned amplifier, (b) frequency response.*

We wish to build a negative-feedback oscillatory system using "LC-tuned" amplifier stages. Figure 8.16(a) shows such a stage, where C_1 denotes the total capacitance seen at the output node and R_p the equivalent parallel resistance of the tank at the resonance frequency. We neglect C_{GD} here but will see that it can be readily included in the final oscillator topology.

Let us examine the frequency response of the stage. At very low frequencies, L_1 dominates the load and

$$\frac{V_{out}}{V_{in}} \approx -g_m L_1 s. \tag{8.31}$$

That is, $|V_{out}/V_{in}|$ is very small and $\angle(V_{out}/V_{in})$ remains around -90° [Fig. 8.16(b)]. At the resonance frequency, ω_0, the tank reduces to R_p and

$$\frac{V_{out}}{V_{in}} = -g_m R_p. \tag{8.32}$$

The phase shift from the input to the output is thus equal to -180°. At very high frequencies, C_1 dominates, yielding

$$\frac{V_{out}}{V_{in}} \approx -g_m \frac{1}{C_1 s}. \tag{8.33}$$

Thus, $|V_{out}/V_{in}|$ diminishes and $\angle(V_{out}/V_{in})$ approaches $+90^\circ$ $(=-270^\circ)$.

Can the circuit of Fig. 8.16(a) oscillate if its input and output are shorted? As evidenced by the open-loop magnitude and phase plots shown in Fig. 8.16(b), no frequency satisfies Barkhausen's criteria; the total phase shift fails to reach 360° at any frequency.

Upon closer examination, we recognize that the circuit provides a phase shift of 180° with possibly adequate gain ($g_m R_p$) at ω_0. We simply need to increase the phase shift to 360°, perhaps by inserting another stage in the loop. Illustrated in Fig. 8.17(a), the idea is to cascade two identical LC-tuned stages so that, at resonance, the total phase shift around the

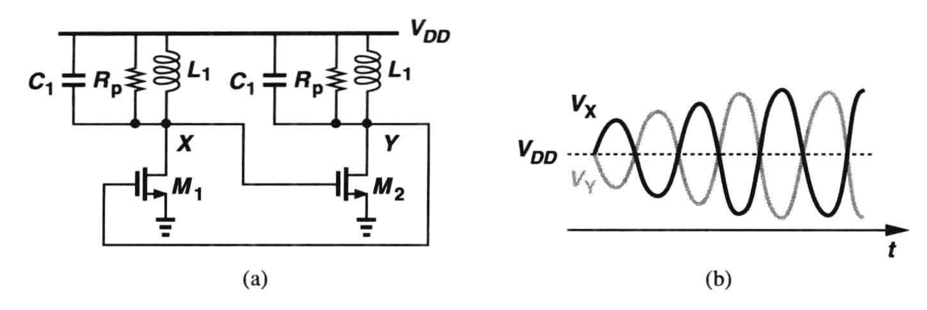


Figure 8.17 *Cascade of two tuned amplifiers in feedback loop.*

loop is equal to 360°. The circuit oscillates if the loop gain is equal to or greater than unity:

$$(g_m R_p)^2 \geq 1. \tag{8.34}$$

Example 8.10

Assuming that the circuit of Fig. 8.17(a) oscillates, plot the voltage waveforms at X and Y.

Solution:

At $t = 0$, $V_X = V_Y = V_{DD}$. As a noise component at ω_0 is amplified and circulated around the loop, V_X and V_Y begin to grow while maintaining a 180° phase difference [Fig. 8.17(b)]. A unique attribute of inductive loads is that they can provide peak voltages above the supply.[1] The growth of V_X and V_Y ceases when M_1 and M_2 enter the triode region for part of the period, reducing the loop gain.

The above circuit can be redrawn as shown in Fig. 8.18(a) and is called a "cross-coupled" oscillator due to the connection of M_1 and M_2. Forming the core of most RF

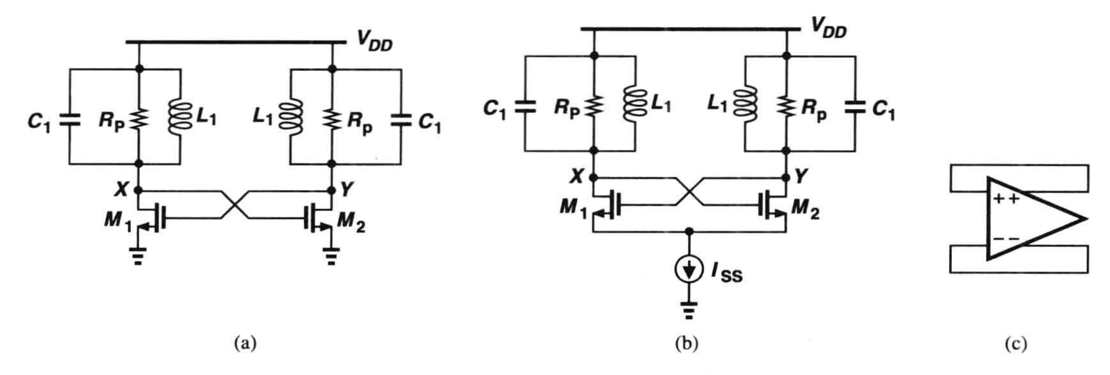


Figure 8.18 *(a) Simple cross-coupled oscillator, (b) addition of tail current, (c) equivalence to a differential pair placed in feedback.*

1. If L_1 has no series resistance, then its average voltage drop must be zero; thus, V_X and V_Y must go above V_{DD} and below V_{DD}.

oscillators used in practice, this topology entails many interesting properties and will be studied from different perspectives in this chapter.

Let us compute the oscillation frequency of the circuit. The capacitance at X includes C_{GS2}, C_{DB1}, and the effect of C_{GD1} and C_{GD2}. We note that (a) C_{GD1} and C_{GD2} are in parallel, and (b) the total voltage change across $C_{GD1} + C_{GD2}$ is equal to twice the voltage change at X (or Y) because V_X and V_Y vary differentially. Thus,

$$\omega_{osc} = \frac{1}{\sqrt{L_1(C_{GS2} + C_{DB1} + 4C_{GD} + C_1)}}. \tag{8.35}$$

Here, C_1 denotes the parasitic capacitance of L_1 plus the input capacitance of the next stage.

The oscillator of Fig. 8.18(a) suffers from poorly-defined bias currents. Since the average V_{GS} of each transistor is equal to V_{DD}, the currents strongly depend on the mobility, threshold voltage, and temperature. With differential V_X and V_Y, we surmise that M_1 and M_2 can operate as a differential pair if they are tied to a tail current source. Shown in Fig. 8.18(b), the resulting circuit is more robust and can be viewed as an inductively-loaded differential pair with positive feedback [Fig. 8.18(c)]. The oscillation amplitude grows until the pair experiences saturation. We sometimes refer to this circuit as the "tail-biased oscillator" to distinguish it from other cross-coupled topologies.

Example 8.11

Compute the voltage swings in the circuit of Fig. 8.18(b) if M_1 and M_2 experience complete current switching with abrupt edges.

Solution:

From Example 8.7, the drain current of each transistor swings between zero and I_{SS}, yielding a peak differential output swing of

$$V_{XY} \approx \frac{4}{\pi} I_{SS} R_p. \tag{8.36}$$

The above-supply swings in the cross-coupled oscillator of Fig. 8.18(b) raise concern with respect to transistor reliability. The instantaneous voltage difference between any two terminals of M_1 or M_2 must remain below the maximum value allowed by the technology. Figure 8.19 shows a "snapshot" of the circuit when M_1 is off and M_2 is on. Each transistor may experience stress under the following conditions: (1) The drain reaches $V_{DD} + V_a$, where V_a is the peak single-ended swing, e.g., $(2/\pi)I_{SS}R_p$, while the gate drops to $V_{DD} - V_a$. The transistor remains off, but its drain-gate voltage is equal to $2V_a$ and its drain-source voltage is greater than $2V_a$ (why?). (2) The drain falls to $V_{DD} - V_a$ while the gate rises to $V_{DD} + V_a$. Thus, the gate-drain voltage reaches $2V_a$ and the gate-source voltage exceeds $2V_a$. We note that both V_{DS1} and V_{GS2} may assume excessively large values. Proper choice of V_a, I_{SS}, and device dimensions avoids stressing the transistors.

The reader may wonder how the inductance value and the device dimensions are selected in the cross-coupled oscillator. We defer the design procedure to after we have studied voltage-controlled oscillators and phase noise (Section 8.8).

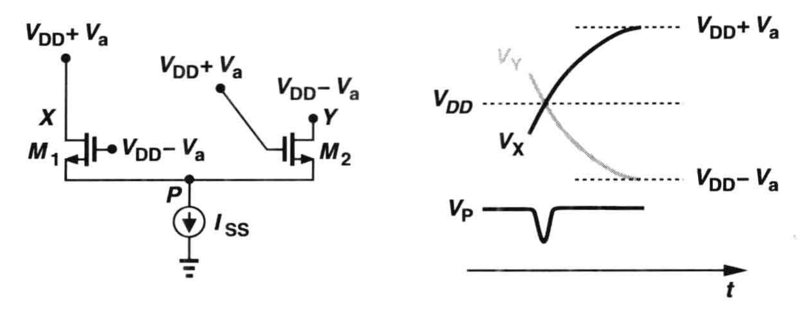


Figure 8.19 *Voltage swings in cross-coupled oscillator.*[2]

Example 8.12

A student claims that the cross-coupled oscillator of Fig. 8.18(b) exhibits no supply sensitivity if the tail current source is ideal. Is this true?

Solution:

No, it is not. The drain-substrate capacitance of each transistor sustains an average voltage equal to V_{DD} (Fig. 8.20). Thus, supply variations modulate this capacitance and hence the oscillation frequency.

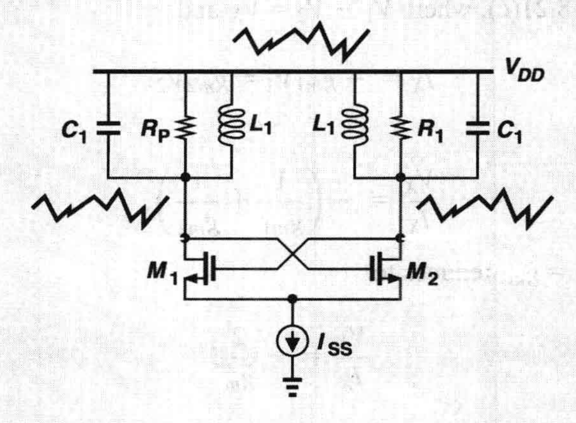


Figure 8.20 *Modulation of drain junction capacitances by V_{DD}.*

While conceived as a feedback system, the cross-coupled oscillator also lends itself to the one-port view described in Section 8.2.2. Let us first redraw the circuit as shown in Fig. 8.21(a) and note that, for small differential waveforms at X and Y, V_N does not change even if it is not connected to V_{DD}. Disconnecting this node from V_{DD} (only for small-signal analysis) and recognizing that the series combination of two identical tanks

2. The voltage at node *P falls* at the crossings of V_X and V_Y if M_1 and M_2 do not enter the triode region at any point. On the other hand, if each transistor enters the deep triode region in a half cycle, then V_P is low most of the time and *rises* at the crossings at V_X and V_Y.

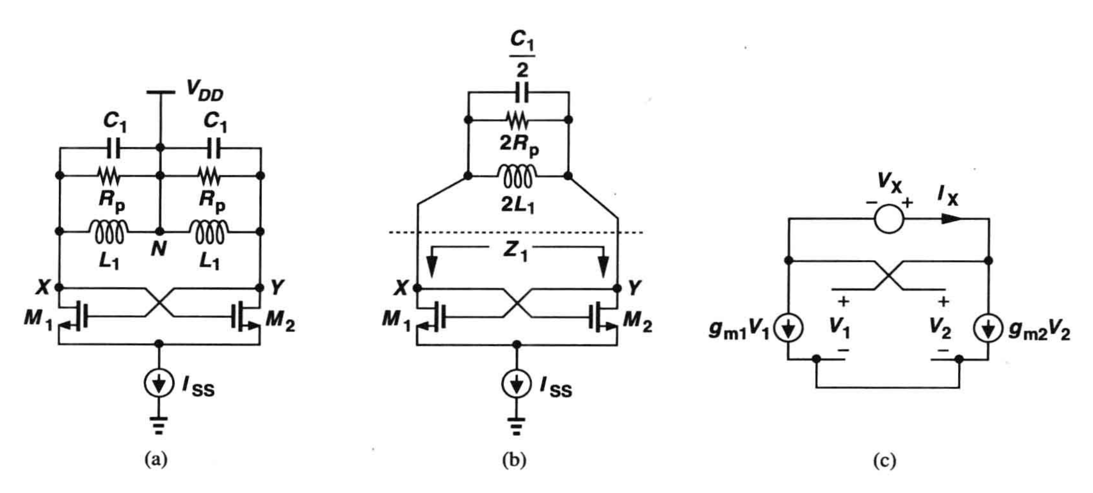


Figure 8.21 *(a) Redrawing of cross-coupled oscillator, (b) load tanks merged, (c) equivalent circuit of cross-coupled pair.*

can be represented as a single tank, we arrive at the circuit depicted in Fig. 8.21(b). We can now view the oscillator as a lossy resonator ($2L_1$, $C_1/2$, and $2R_p$) tied to the port of an active circuit (M_1, M_2, and I_{SS}), expecting that the latter replenishes the energy lost in the former. That is, Z_1 must contain a negative resistance. This can be seen from the equivalent circuit shown in Fig. 8.21(c), where $V_1 - V_2 = V_X$ and

$$I_X = -g_{m1}V_1 = g_{m2}V_2. \tag{8.37}$$

It follows that

$$\frac{V_X}{I_X} = -\left(\frac{1}{g_{m1}} + \frac{1}{g_{m2}}\right), \tag{8.38}$$

which, for $g_{m1} = g_{m2} = g_m$, reduces to

$$\frac{V_X}{I_X} = -\frac{2}{g_m}. \tag{8.39}$$

For oscillation to occur, the negative resistance must cancel the loss of the tank:

$$\frac{2}{g_m} \leq 2R_p \tag{8.40}$$

and hence

$$g_m R_p \geq 1. \tag{8.41}$$

As expected, this condition is identical to that expressed by Eq. (8.34).[3]

3. This topology is also called a "negative-G_m oscillator." This is not quite correct because it does not contain a negative *transconductance* but a negative *conductance*.

Choice of g_m The foregoing studies may suggest that the g_m of the cross-coupled transistors in Fig. 8.18(b) can be chosen slightly greater than R_p of the tank to ensure oscillation. However, this choice leads to small voltage swings; if the swings are large, e.g., if M_1 and M_2 switch completely, then the g_m falls below $1/R_p$ for part of the period, failing to sustain oscillation. (That is, with $g_m \approx 1/R_p$, M_1 and M_2 must remain linear to avoid compression.) In practice, therefore, we design the circuit for nearly complete current steering between M_1 and M_2, inevitably choosing a g_m quite higher than $1/R_p$.

8.4 THREE-POINT OSCILLATORS

As observed in Section 8.2.2, the circuit of Fig. 8.14(a) can be attached to an inductor so as to form an oscillator. Note that the derivation of the impedance Z_{in} does not assume any terminal is grounded. Thus, three different oscillator topologies can be obtained by grounding each of the transistor terminals. Figures 8.22(a), (b), and (c) depict the resulting circuits if the source, the gate, or the drain is (ac) grounded, respectively. In each case, a current source defines the bias current of the transistor. [The gate of M_1 in Fig. 8.22(b) and the left terminal of L_1 in Fig. 8.22(c) must be tied to a proper potential, e.g., $V_b - V_{DD}$.]

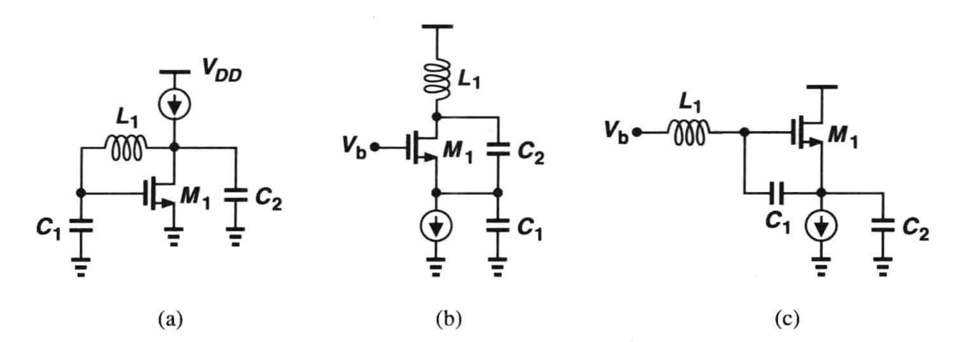


Figure 8.22 *Variants of three-point oscillator, (a) with source grounded, (b) with gate grounded (Colpitts oscillator), (c) with drain grounded (Clapp oscillator).*

It is important to bear in mind that the operation frequency and startup condition of all three oscillators in Fig. 8.22 are given by Eqs. (8.26) and (8.30), respectively. Specifically, the transistor must provide sufficient transconductance to satisfy

$$g_m R_p \geq 4 \tag{8.42}$$

if $C_1 = C_2$. This condition is more stringent than Eq. (8.34) for the cross-coupled oscillator, suggesting that the circuits of Fig. 8.22 may fail to oscillate if the inductor Q is not very high. This is the principal disadvantage of these oscillators and the reason for their lack of popularity.

Another drawback of the circuits shown in Fig. 8.22 is that they produce only single-ended outputs. It is possible to couple two copies of one oscillator so that they operate differentially. Shown in Fig. 8.23 is an example, where two instances of the oscillator in Fig. 8.22(c) are coupled at node P. Resistor R_1 establishes a dc level equal to V_{DD} at P and at the gates of M_1 and M_2. More importantly, if chosen properly, this resistor

prohibits common-mode oscillation. To understand this point, suppose X and Y swing in phase and so do A and B, creating in-phase currents through L_1 and L_2. The two half circuits then collapse into one, and R_1 appears in *series* with the parallel combination of L_1 and L_2, thereby lowering their Q. No CM oscillation can occur if R_1 is sufficiently large. For differential waveforms, on the other hand, L_1 and L_2 carry equal and opposite currents, forcing P to ac ground.

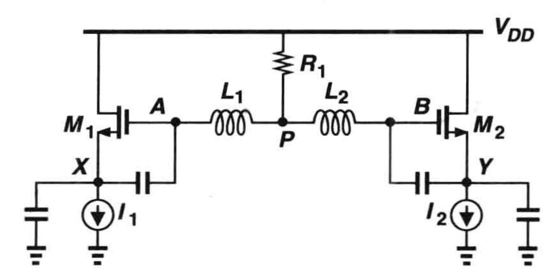


Figure 8.23 *Differential version of a three-point oscillator.*

Even with differential outputs, the circuit of Fig. 8.23 may be inferior to the cross-coupled oscillator of Fig. 8.18(b)—not only for the more stringent startup condition but also because the noise of I_1 and I_2 directly corrupts the oscillation. The circuit nonetheless has been used in some designs.

8.5 VOLTAGE-CONTROLLED OSCILLATORS

Most oscillators must be tuned over a certain frequency range. We therefore wish to construct oscillators whose frequency can be varied electronically. "Voltage-controlled oscillators" (VCOs) are an example of such circuits.

Figure 8.24 conceptually shows the desired behavior of a VCO. The output frequency varies from ω_1 to ω_2 (the required tuning range) as the control voltage, V_{cont}, goes from V_1 to V_2. The slope of the characteristic, K_{VCO}, is called the "gain" or "sensitivity" of the VCO and expressed in rad/Hz/V. We formulate this characteristic as

$$\omega_{out} = K_{VCO}V_{cont} + \omega_0, \tag{8.43}$$

where ω_0 denotes the intercept point on the vertical axis. As explained in Chapter 9, it is desirable that this characteristic be relatively linear, i.e., K_{VCO} not change significantly across the tuning range.

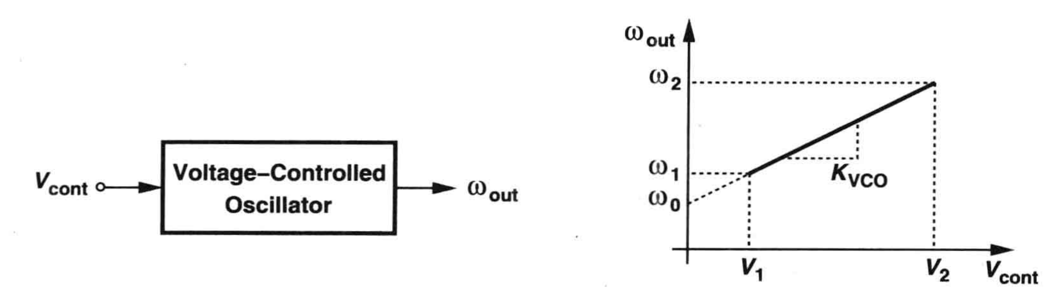


Figure 8.24 *VCO characteristic.*

Example 8.13

As explained in Example 8.12, the cross-coupled oscillator exhibits sensitivity to V_{DD}. Considering V_{DD} as the "control voltage," determine the gain.

Solution:

We have

$$\omega_{osc} = \frac{1}{\sqrt{L_1(C_1 + C_{DB})}}, \tag{8.44}$$

where C_1 includes all circuit capacitances except C_{DB}. Thus,

$$K_{VCO} = \frac{\partial \omega_{out}}{\partial V_{DD}} \tag{8.45}$$

$$= \frac{\partial \omega_{osc}}{\partial C_{DB}} \cdot \frac{\partial C_{DB}}{\partial V_{DD}}. \tag{8.46}$$

The junction capacitance, C_{DB}, is approximated as

$$C_{DB} = \frac{C_{DB0}}{\left(1 + \dfrac{V_{DD}}{\phi_B}\right)^m}, \tag{8.47}$$

where ϕ_B denotes the junction's built-in potential and m is around 0.3 to 0.4. It follows from Eqs. (8.46) and (8.47) that

$$K_{VCO} = \frac{-1}{2\sqrt{L_1}} \cdot \frac{1}{\sqrt{C_1 + C_{DB}}(C_1 + C_{DB})} \cdot \frac{-mC_{DB0}}{\phi_B\left(1 + \dfrac{V_{DD}}{\phi_B}\right)^{m+1}} \tag{8.48}$$

$$= \frac{C_{DB}}{C_1 + C_{DB}} \cdot \frac{m}{2\phi_B + 2V_{DD}} \omega_{osc}. \tag{8.49}$$

In order to vary the frequency of an LC oscillator, the resonance frequency of its tank(s) must be varied. Since it is difficult to vary the inductance electronically, we only vary the capacitance by means of a varactor. As explained in Chapter 7, MOS varactors are more commonly used than *pn* junctions, especially in low-voltage design. We thus construct our first VCO as shown in Fig. 8.25(a), where varactors M_{V1} and M_{V2} appear in *parallel* with the tanks (if V_{cont} is provided by an ideal voltage source). Note that the gates of the varactors are tied to the oscillator nodes and the source/drain/n-well terminals to V_{cont}. This avoids loading X and Y with the capacitance between the n-well and the substrate.

Since the gates of M_{V1} and M_{V2} reside at an average level equal to V_{DD}, their gate-source voltage remains positive and their capacitance *decreases* as V_{cont} goes from zero to V_{DD} [Fig. 8.25(b)]. This behavior persists even in the presence of large voltage swings at X and Y and hence across M_{V1} and M_{V2}. The key point here is that the *average* voltage

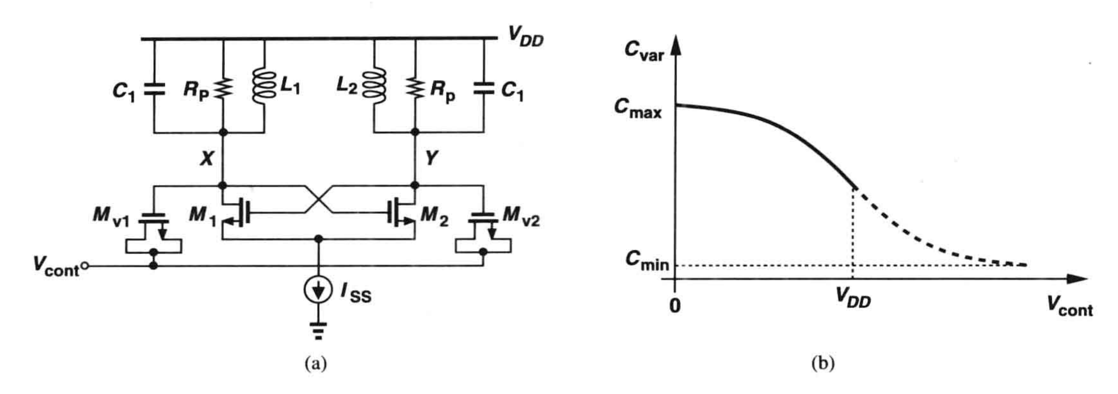


Figure 8.25 *(a) VCO using MOS varactors, (b) range of varactor capacitance used in (a).*

across each varactor varies from V_{DD} to zero as V_{cont} goes from zero to V_{DD}, thus creating a monotonic decrease in their capacitance. The oscillation frequency can thus be expressed as

$$\omega_{osc} = \frac{1}{\sqrt{L_1(C_1 + C_{var})}}, \tag{8.50}$$

where C_{var} denotes the average value of each varactor's capacitance.

The reader may wonder why capacitors C_1 have been included in the oscillator of Fig. 8.25(a). It appears that, without C_1, the varactors can vary the frequency to a greater extent, thereby providing a wider tuning range. This is indeed true, and we rarely need to add a constant capacitance to the tank deliberately. In other words, C_1 simply models the inevitable capacitances appearing at X and Y: (1) C_{GS}, C_{GD} (with a Miller multiplication factor of two), and C_{DB} of M_1 and M_2, (2) the parasitic capacitance of each inductor, and (3) the input capacitance of the next stage (e.g., a buffer, divider, or mixer). As mentioned in Chapter 4, the last component becomes particularly significant on the transmit side due to the "propagation" of the capacitance from the input of the PA to the input of the upconversion mixers.

The above VCO topology merits two remarks. First, the varactors are stressed for part of the period if V_{cont} is near ground and V_X (or V_Y) rises significantly above V_{DD}. Second, as depicted in Fig. 8.25(b), only about half of $C_{max} - C_{min}$ is utilized in the tuning. We address these issues later in this chapter.

As explained in Chapter 7, symmetric spiral inductors excited by differential waveforms exhibit a higher Q than their single-ended counterparts. For this reason, L_1 and L_2 in Fig. 8.25 are typically realized as a single symmetric structure. Figure 8.26 illustrates the idea and its circuit representation. The point of symmetry of the inductor (its "center tap") is tied to V_{DD}. In some of our analyses, we omit the center tap connection for the sake of simplicity.

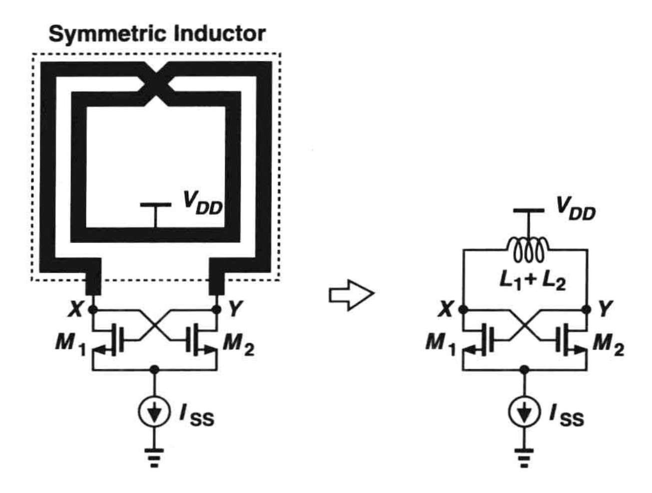


Figure 8.26 *Oscillator using symmetric inductor.*

Example 8.14

The symmetric inductor in Fig. 8.26 has a value of 2 nH and a Q of 10 at 10 GHz. What is the minimum required transconductance of M_1 and M_2 to guarantee startup?

Solution:

The parallel equivalent resistance of $L_1 + L_2 = 2$ nH is equal to $Q(L_1 + L_2)\omega = 1.26$ kΩ. From Eq. (8.40), we have

$$g_{m1,2} \geq (630\ \Omega)^{-1}. \tag{8.51}$$

Alternatively, we can decompose the inductor into L_1 and L_2 and return to the circuit of Fig. 8.18(b). In this case, $R_p = QL_1\omega = QL_2\omega = 630\ \Omega$, and $g_{m1,2}R_p \geq 1$. Thus, $g_{m1,2} \geq (630\ \Omega)^{-1}$. The point here is that, for frequency and startup calculations, we can employ the one-port model with $L_1 + L_2$ as one inductor or the feedback model with L_1 and L_2 belonging to two stages.

The VCO of Fig. 8.25(a) provides an output CM level near V_{DD}, an advantage or disadvantage depending on the next stage (Section 8.9).

8.5.1 Tuning Range Limitations

While a robust, versatile topology, the cross-coupled VCO of Fig. 8.25(a) suffers from a narrow tuning range. As mentioned above, the three components comprising C_1 tend to limit the effect of the varactor capacitance variation. Since in (8.50), C_{var} tends to be a

small fraction of the total capacitance, we make a crude approximation, $C_{var} \ll C_1$, and rewrite (8.50) as

$$\omega_{osc} \approx \frac{1}{\sqrt{L_1 C_1}} \left(1 - \frac{C_{var}}{2C_1}\right). \tag{8.52}$$

If the varactor capacitance varies from C_{var1} to C_{var2}, then the tuning range is given by

$$\Delta\omega_{osc} \approx \frac{1}{\sqrt{L_1 C_1}} \frac{C_{var2} - C_{var1}}{2C_1}. \tag{8.53}$$

For example, if $C_{var2} - C_{var1} = 20\% C_1$, then the tuning range is about $\pm 5\%$ around the center frequency.

What limits the capacitance range of the varactor, $C_{var2} - C_{var1}$? We note from Chapter 7 that $C_{var2} - C_{var1}$ trades with the Q of the varactor: a longer channel reduces the relative contribution of the gate-drain and gate-source overlap capacitances, widening the range but lowering the Q. Thus, the tuning range trades with the overall tank Q (and hence with the phase noise).

Another limitation on $C_{var2} - C_{var1}$ arises from the available range for the control voltage of the oscillator, V_{cont} in Fig. 8.25(a). This voltage is generated by a "charge pump" (Chapter 9), which, as any other analog circuit, suffers from a limited output voltage range. For example, a charge pump running from a 1-V supply may not be able to generate an output below 0.1 V or above 0.9 V. The characteristic of Fig. 8.25(b) therefore reduces to that depicted in Fig. 8.27.

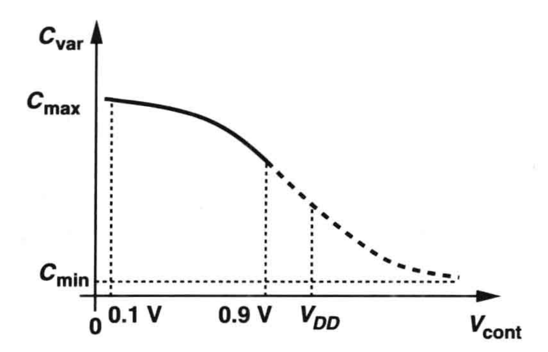


Figure 8.27 *Varactor range used with input limited between 0.1 V and 0.9 V.*

The foregoing tuning limitations prove serious in LC VCO design. We introduce in Section 8.6 a number of oscillator topologies that provide a wider tuning range—but at the cost of other aspects of the performance.

8.5.2 Effect of Varactor Q

As observed in the previous section, the varactor capacitance is but a small fraction of the total tank capacitance. We therefore surmise that the resistive loss of the varactor lowers the overall Q of the tank only to some extent. Let us begin with a fundamental observation.

Example 8.15

A lossy inductor and a lossy capacitor form a parallel tank. Determine the overall Q in terms of the quality factor of each.

Solution:

The loss of an inductor or a capacitor can be modeled by a parallel resistance (for a narrow frequency range). We therefore construct the tank as shown in Fig. 8.28, where the

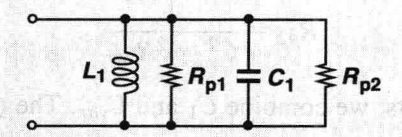


Figure 8.28 *Tank consisting of lossy inductor and lossy capacitor.*

inductor and capacitor Q's are respectively given by

$$Q_L = \frac{R_{p1}}{L_1\omega} \tag{8.54}$$

$$Q_C = R_{p2}C_1\omega. \tag{8.55}$$

Note that, in the vicinity of resonance, $L_1\omega = (C_1\omega)^{-1}$. Merging R_{p1} and R_{p2} yields the overall Q:

$$Q_{tot} = \frac{R_{p1}R_{p2}}{R_{p1}+R_{p2}} \cdot \frac{1}{L_1\omega} \tag{8.56}$$

$$= \frac{1}{\frac{L_1\omega}{R_{p1}} + \frac{L_1\omega}{R_{p2}}} \tag{8.57}$$

$$= \frac{1}{\frac{L_1\omega}{R_{p1}} + \frac{1}{R_{p2}C_1\omega}}. \tag{8.58}$$

It follows that

$$\frac{1}{Q_{tot}} = \frac{1}{Q_L} + \frac{1}{Q_C}. \tag{8.59}$$

To quantify the effect of varactor loss, consider the tank shown in Fig. 8.29(a), where R_{p1} models the loss of the inductor and R_{var} the equivalent series resistance of the varactor. We wish to compute the Q of the tank. Transforming the series combination of C_{var} and

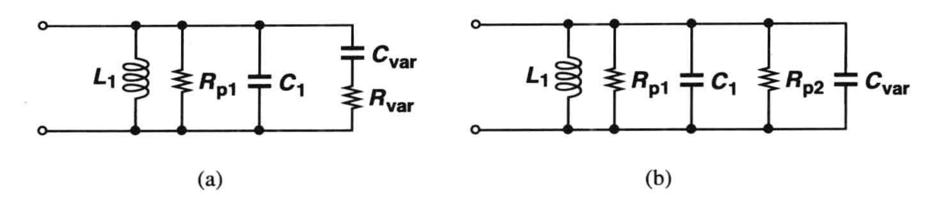


Figure 8.29 *(a) Tank using lossy varactor, (b) equivalent circuit.*

R_{var} to a parallel combination [Fig. 8.29(b)], we have from Chapter 2

$$R_{p2} = \frac{1}{C_{var}^2\omega^2 R_{var}}. \tag{8.60}$$

To utilize our previous results, we combine C_1 and C_{var}. The Q associated with $C_1 + C_{var}$ is equal to

$$Q_C = R_{P2}(C_1 + C_{var})\omega \tag{8.61}$$

$$= \frac{C_1 + C_{var}}{C_{var}^2 \omega R_{var}}. \tag{8.62}$$

Recognizing that $Q_{var} = (C_{var}\omega R_{var})^{-1}$, we have

$$Q_C = \left(1 + \frac{C_1}{C_{var}}\right) Q_{var}. \tag{8.63}$$

In other words, the Q of the varactor is "boosted" by a factor of $1 + C_1/C_{var}$. The overall tank Q is therefore given by

$$\frac{1}{Q_{tot}} = \frac{1}{Q_L} + \frac{1}{\left(1 + \frac{C_1}{C_{var}}\right) Q_{var}}. \tag{8.64}$$

For frequencies as high as several tens of gigahertz, the first term in Eq. (8.64) is dominant (unless a long channel is chosen for the varactors).

Equation (8.64) can be generalized if the tank consists of an ideal capacitor, C_1, and lossy capacitors, C_2-C_n, that exhibit a series resistance of R_2-R_n, respectively. The reader can prove that

$$\frac{1}{Q_{tot}} = \frac{1}{Q_L} + \frac{C_2}{C_{tot}}\frac{1}{Q_2} + \cdots + \frac{C_n}{C_{tot}}\frac{1}{Q_n}, \tag{8.65}$$

where $C_{tot} = C_1 + \cdots + C_n$ and $Q_j = (R_j C_j \omega)^{-1}$.

8.6 LC VCOs WITH WIDE TUNING RANGE

8.6.1 VCOs with Continuous Tuning

The tuning range obtained from the C–V characteristic depicted in Fig. 8.27 may prove prohibitively narrow, particularly because the capacitance range corresponding to *negative*

V_{GS} (for $V_{cont} > V_{DD}$) remains unused. We must therefore seek oscillator topologies that allow both positive and negative (average) voltages across the varactors, utilizing almost the entire range from C_{min} to C_{max}.

Figure 8.30(a) shows one such topology. Unlike the tail-biased configuration studied in Section 8.3, this circuit defines the bias currents of M_1 and M_2 by a *top* current source, I_{DD}. We analyze this circuit by first computing the output common-mode level. In the absence of oscillation, the circuit reduces to that shown in Fig. 8.30(b), where M_1 and M_2 share I_{DD} equally and are configured as diode-connected devices. Thus, the CM level is simply given by the gate-source voltage of a diode-connected transistor carrying a current of $I_{DD}/2$.[4] For example, for square-law devices,

$$V_{GS1,2} = \sqrt{\frac{I_{DD}}{\mu_n C_{ox}(W/L)}} + V_{TH}. \tag{8.66}$$

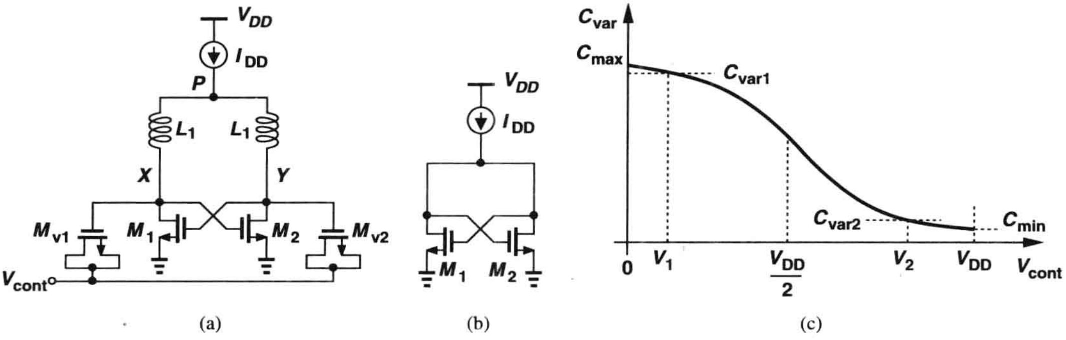


Figure 8.30 *(a) Top-biased VCO, (b) equivalent circuit for CM calculation, (c) varactor range used.*

We select the transistor dimensions such that the CM level is approximately equal to $V_{DD}/2$. Consequently, as V_{cont} varies from 0 to V_{DD}, the gate-source voltage of the varactors, $V_{GS,var}$, goes from $+V_{DD}/2$ to $-V_{DD}/2$, sweeping almost the entire capacitance range from C_{min} to C_{max} [Fig. 8.30(c)]. In practice, the circuit producing V_{cont} (the charge pump) can handle only the voltage range from V_1 to V_2, yielding a capacitance range from C_{var1} to C_{var2}.

The startup condition, oscillation frequency, and output swing of the oscillator shown in Fig. 8.30(a) are similar to those derived for the tail-biased circuit of Fig. 8.18(b). Also, L_1 and L_2 are realized as a single symmetric inductor so as to achieve a higher Q; the center tap of the inductor is tied to I_{DD}.

While providing a wider range than its tail-biased counterpart, the topology of Fig. 8.30(a) suffers from a higher phase noise. As studied in Section 8.7, this penalty arises primarily from the modulation of the output CM level (and hence the varactors) by the noise current of I_{DD}, as evidenced by Eq. (8.66). This effect does not occur in the tail-biased oscillator because the output CM level is "pinned" at V_{DD} by the low dc resistance of the inductors. The following example illustrates this difference.

4. With large-signal oscillation, the nonlinearity of M_1 and M_2 shifts the output CM level slightly, but we neglect this effect here.

Example 8.16

The tail or top bias current in the above oscillators is changed by ΔI. Determine the change in the voltage across the varactors.

Solution:

As shown in the tail-biased topology of Fig. 8.31(a), each inductor contains a small low-frequency resistance, r_s (typically no more than 10 Ω). If I_{SS} changes by ΔI, the output CM level changes by $\Delta V_{CM} = (\Delta I/2)r_s$, and so does the voltage across each varactor. In the top-biased circuit of Fig. 8.31(b), on the other hand, a change of ΔI flows through two diode-connected transistors, producing an output CM change of $\Delta V_{CM} = (\Delta I/2)(1/g_m)$. Since $1/g_m$ is typically in the range of a few hundred ohms, the top-biased topology suffers from a much higher varactor voltage modulation.

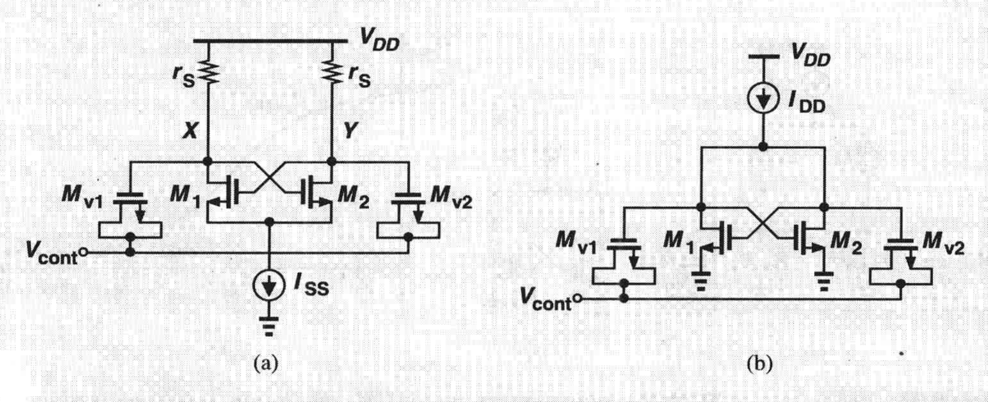


Figure 8.31 *Output CM dependence on bias current in (a) tail-biased and (b) top-biased VCOs.*

Example 8.17

What is the change in the oscillation frequency in the above example?

Solution:

Since a CM change at X and Y is indistinguishable from a change in V_{cont}, we have

$$\Delta\omega = K_{VCO}\Delta V_{CM} \tag{8.67}$$

$$= K_{VCO}\frac{\Delta I}{2}r_s \quad \text{or} \quad K_{VCO}\frac{\Delta I}{2}\frac{1}{g_m}. \tag{8.68}$$

In order to avoid varactor modulation due to the noise of the bias current source, we return to the tail-biased topology but employ *ac coupling* between the varactors and the core so as to allow positive and negative voltages across the varactors. Illustrated in Fig. 8.32(a),

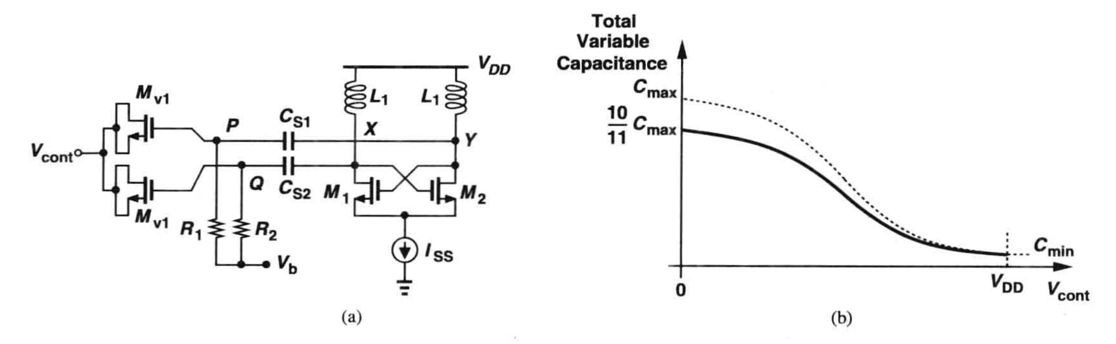


Figure 8.32 *(a) VCO using capacitor coupling to varactors, (b) reduction of tuning range as a result of finite C_{S1} and C_{S2}.*

the idea is to define the dc voltage at the gate of the varactors by V_b ($\approx V_{DD}/2$) rather than V_{DD}. Thus, in a manner similar to that shown in Fig. 8.30(c), the voltage across each varactor goes from $-V_{DD}/2$ to $+V_{DD}/2$ as V_{cont} varies from 0 to V_{DD}, maximizing the tuning range.

The principal drawback of the above circuit stems from the parasitics of the coupling capacitors. In Fig. 8.32(a), C_{S1} and C_{S2} must be *much greater* than the maximum capacitance of the varactors, C_{max}, so that the capacitance range presented by the varactors to the tanks does not shrink substantially. If $C_{S1} = C_{S2} = C_S$, then in Eq. (8.53), C_{var2} and C_{var1} must be placed in series with C_S, yielding

$$\Delta\omega_{os} \approx \frac{1}{\sqrt{L_1C_1}} \cdot \frac{1}{2C_1} \cdot \frac{C_S^2(C_{var2} - C_{var1})}{(C_S + C_{var2})(C_S + C_{var1})}. \tag{8.69}$$

For example, if $C_S = 10C_{max}$, then the series combination yields a maximum capacitance of $(10C_{max} \cdot C_{max})/(11C_{max}) = (10/11)C_{max}$, i.e., about 10% less than C_{max}. Thus, as shown in Fig. 8.32(b), the capacitance range decreases by about 10%. Equivalently, the maximum-to-minimum capacitance ratio falls from C_{max}/C_{min} to $(10C_{max} + C_{min})/(11C_{min}) \approx (10/11)(C_{max}/C_{min})$.

The choice of $C_S = 10C_{max}$ reduces the capacitance range by 10% but introduces substantial parasitic capacitances at X and Y or at P and Q. This is because integrated capacitors suffer from parasitic capacitances to the substrate. An example is depicted in Fig. 8.33(a), where a sandwich of metal layers from metal 6 to metal 9 forms the wanted capacitance between nodes A and B, and the capacitance between the bottom layer and the substrate, C_b, appears from node A to ground. We must therefore choose the number of layers in the sandwich so as to minimize C_b/C_{AB}. Plotted in Fig. 8.33(b) is this ratio as a function of the number of the layers, assuming that we begin with the top layers. For only metal 9 and metal 8, C_b is small, but so is C_{AB}. As more layers are stacked, C_b increases more slowly than C_{AB} does, yielding a declining ratio. As the bottom layer approaches the substrate, C_b begins to increase more rapidly than C_{AB}, producing the minimum shown in Fig. 8.33(b). In other words, C_b/C_{AB} typically exceeds 5%.

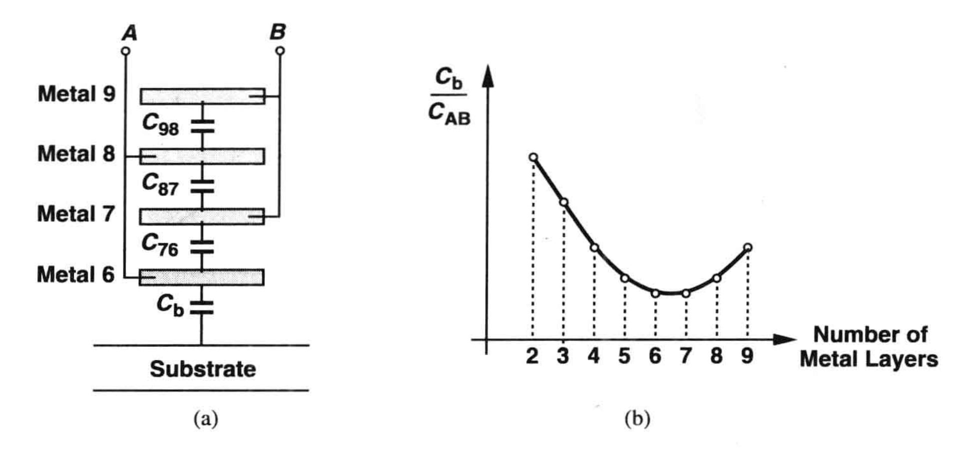


Figure 8.33 *(a) Capacitor realized as parallel metal plates, (b) relative parasitic capacitance as a function of the number of metal layers.*

Let us now study the effect of the parasitics of C_{S1} and C_{S2} in Fig. 8.32(a). From Eq. (8.53), we note that a larger C_1 further limits the tuning range. In other words, the numerator of (8.53) *decreases* due to the series effect of C_S, and the denominator of (8.53) *increases* due to the parasitic capacitance of C_S. To formulate these limitations, we assume a typical case, $C_{max} \approx 2C_{min}$, and also $C_{var2} \approx C_{max}$, $C_{var1} \approx C_{min}$, $C_S = 10C_{max}$, and $C_b = 0.05C_S = 0.5C_{max}$. Equation (8.69) thus reduces to

$$\Delta\omega_{osc} \approx \frac{1}{\sqrt{L_1(C_1 + 0.5C_{max})}} \times \frac{1}{2(C_1 + 0.5C_{max})} \times$$

$$\times \frac{C_S^2(C_{max} - 0.5C_{max})}{(10C_{max} + C_{max})(10C_{max} + 0.5C_{max})} \tag{8.70}$$

$$\approx \frac{1}{\sqrt{L_1(C_1 + 0.5C_{max})}} \times \frac{0.43C_{max}}{2(C_1 + 0.5C_{max})}. \tag{8.71}$$

Example 8.18

The VCO of Fig. 8.32(a) is designed for a tuning range of 10% without the series effect of C_S and parallel effect of C_b. If $C_S = 10C_{max}$, $C_{max} = 2C_{min}$, and $C_b = 0.05C_S$, determine the actual tuning range.

Solution:

Without the effects of C_S and C_b, Eq. (8.53) applies:

$$\Delta\omega_{osc} \approx \frac{1}{\sqrt{L_1C_1}} \frac{0.5C_{max}}{2C_1}. \tag{8.72}$$

Example 8.18 *(Continued)*

For this range to reach 10% of the center frequency, we have

$$C_{max} = \frac{2}{5}C_1. \tag{8.73}$$

With the effects of C_S and C_b, Eq. (8.71) yields

$$\Delta\omega_{osc} \approx \frac{1}{\sqrt{L_1(1.2C_1)}} \times \frac{0.43}{6} \tag{8.74}$$

$$\approx \frac{7.2\%}{\sqrt{1.2L_1C_1}}. \tag{8.75}$$

The tuning range therefore falls to 7.2% (around $\sqrt{1.2L_1C_1}^{-1}$).

In the above study, we have assumed that C_b appears at nodes X and Y in Fig. 8.32(a). Alternatively, C_b can be placed at nodes P and Q. We study this case in Problem 8.8, arriving at similar limitations in the tuning range.

A capacitor structure that exhibits lower parasitics than the metal sandwich geometry of Fig. 8.33(a) is shown in Fig. 8.34(a). Called a "fringe" or "lateral-field" capacitor, this topology incorporates closely-spaced narrow metal lines to maximize the fringe capacitance between them. The capacitance per unit volume is larger than that of the metal sandwich, leading to a smaller parasitic.

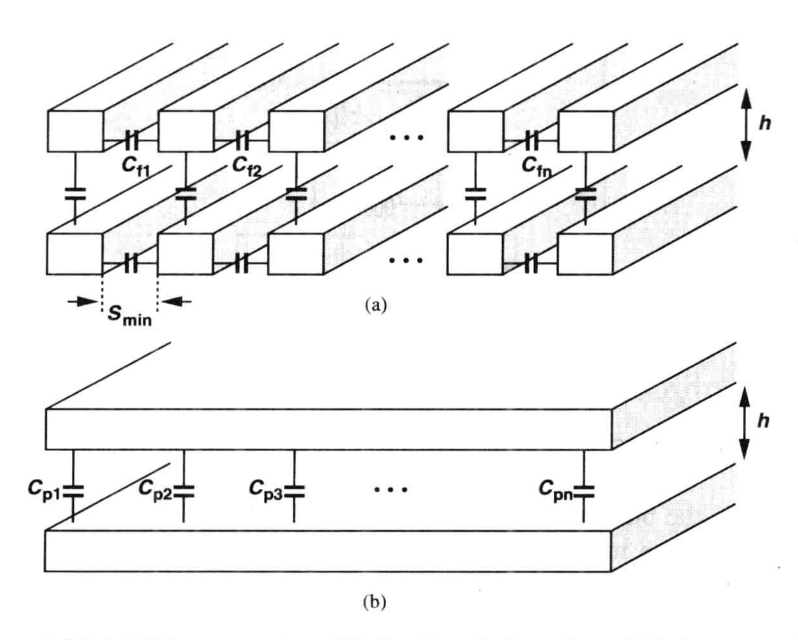


Figure 8.34 *(a) Fringe capacitor, (b) distributed view of parallel-plate capacitor.*

Example 8.19

Explain why the fringe structure provides a larger capacitance per unit volume.

Solution:

Suppose a two-layer metal sandwich capacitor is viewed as the sum of small units [Fig. 8.34(b)]. Given by the vertical spacing, h, the capacitance between the two plates is equal to $C_{p1} + \cdots + C_{pn}$. Now, we decompose each plate into a number of narrow lines, with a spacing equal to the minimum allowed by the technology, S_{min} [Fig. 8.34(a)]. For example, $S_{min} = 0.15$ μm whereas $h = 0.5$ μm. We now recognize that some of the parallel-plate capacitances, C_{pj}, are omitted, but about *twice* as many fringe capacitors, C_{fj}, have been added. Also, since $S_{min} < h$, $C_{fj} > C_{pj}$. Thus, the overall capacitance rises substantially.

Three other issues in Fig. 8.32(a) merit consideration. First, since R_1 and R_2 appear approximately in parallel with the tanks, their value must be chosen much greater than R_p. (Even a tenfold ratio proves inadequate as it lowers the Q by about 10%.) Second, noise on the mid-supply bias, V_b, directly modulates the varactors and must therefore be minimized. Third, as studied in the transceiver design example of Chapter 13, the noise of R_1 and R_2 modulates the varactors, producing substantial phase noise.

Another VCO topology that naturally provides an output CM level approximately equal to $V_{DD}/2$ is shown in Fig. 8.35. The circuit can be viewed as two back-to-back CMOS inverters, except that the sources of the NMOS devices are tied to a tail current, or as a cross-coupled NMOS pair and a cross-coupled PMOS pair sharing the same bias current. Proper choice of device dimensions and I_{SS} can yield a CM level at X and Y around $V_{DD}/2$, thereby maximizing the tuning range.

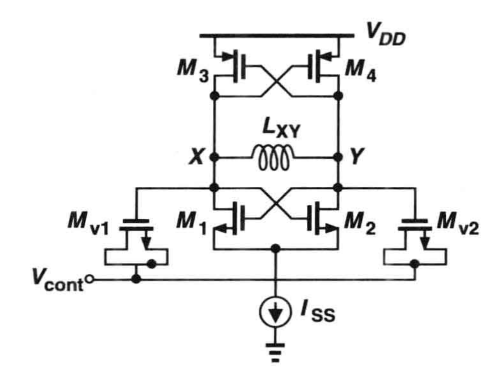


Figure 8.35 *VCO using NMOS and PMOS cross-coupled pairs.*

In this circuit, the bias current is "reused" by the PMOS devices, providing a higher transconductance. But a more important advantage of the above topology over those in Figs. 8.25(a), 8.30(a), and 8.32(a) is that it produces *twice* the voltage swing for a given bias current and inductor design. To understand this point, we assume L_{XY} in the complementary topology is equal to $L_1 + L_2$ in the previous circuits. Thus, L_{XY} presents an equivalent parallel resistance of $2R_p$. Drawing the circuit for each half cycle as shown in Fig. 8.36, we

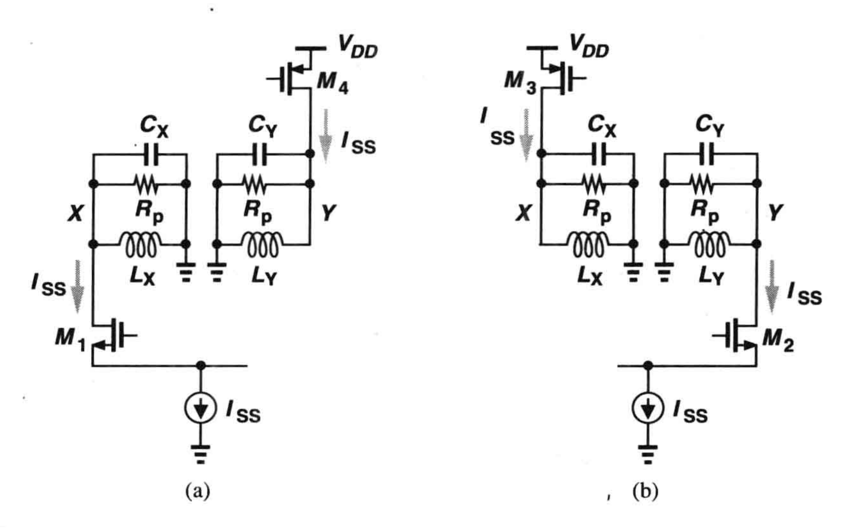


Figure 8.36 *Current flow through floating resonator when (a) M_1 and M_4 are on, and (b) M_2 and M_3 are on.*

recognize that the current in each tank swings between $+I_{SS}$ and $-I_{SS}$, whereas in previous topologies it swings between I_{SS} and zero. The output voltage swing is therefore doubled.

The circuit of Fig. 8.35 nonetheless suffers from two drawbacks. First, for $|V_{GS3}| + V_{GS1} + V_{ISS}$ to be equal to V_{DD}, the PMOS transistors must typically be quite wide, contributing significant capacitance and limiting the tuning range. This is particularly troublesome at very high frequencies, requiring a small inductor and diminishing the above swing advantage. Second, as in the circuit of Fig. 8.30(a), the noise current of the bias current source modulates the output CM level and hence the capacitance of the varactors, producing frequency and phase noise. Following Example 8.17, the reader can show that a change of ΔI in I_{SS} results in a change of $(\Delta I/2)/g_{m3,4}$ in the voltage across each varactor and hence a frequency change of $K_{VCO}(\Delta I/2)/g_{m3,4}$. Owing to the small headroom available for I_{SS}, the noise current of I_{SS}, given by $4kT\gamma g_m$, tends to be large.

Example 8.20

A student attempts to remove the noise of the tail current source by simply eliminating it. Explain the pros and cons of such a topology (Fig. 8.37).

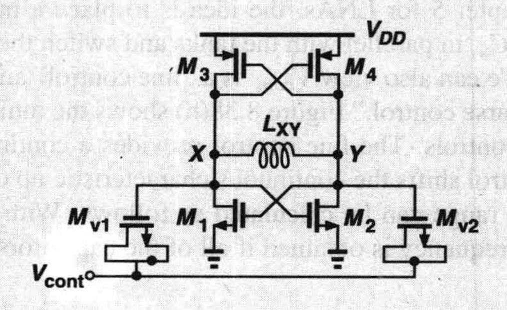


Figure 8.37 *VCO without bias current source.*

(Continues)

Example 8.20 *(Continued)*

Solution:

The circuit indeed avoids frequency modulation due to the tail current noise. Moreover, it saves the voltage headroom associated with the tail current source. However, the circuit is now very sensitive to the supply voltage. For example, a voltage regulator providing V_{DD} may exhibit significant flicker noise, thus modulating the frequency (by modulating the CM level). Furthermore, the bias current of the circuit varies considerably with process and temperature.

8.6.2 Amplitude Variation with Frequency Tuning

In addition to the narrow varactor capacitance range, another factor that limits the *useful* tuning range is the variation of the oscillation amplitude. As the capacitance attached to the tank increases, the amplitude tends to decrease. To formulate this effect, suppose the tank inductor exhibits only a *series* resistance, R_S, (due to metal resistance and skin effect). Recall from Chapter 2 that, for a narrow frequency range and a Q greater than 3,

$$Q = \frac{L_1\omega}{R_S} = \frac{R_p}{L_1\omega} \tag{8.76}$$

and hence

$$R_p = \frac{L_1^2\omega^2}{R_S}. \tag{8.77}$$

Thus, R_p falls in proportion to ω^2 as more capacitance is presented to the tank.[5] For example, a 10% change in ω yields a 20% change in the amplitude.

8.6.3 Discrete Tuning

Our study of varactor tuning in Example 8.18 points to a relatively narrow range. The use of large varactors leads to a high K_{VCO}, making the circuit sensitive to noise on the control voltage. In applications where a substantially wider tuning range is necessary, "discrete tuning" may be added to the VCO so as to achieve a capacitance range well beyond C_{max}/C_{min} of varactors. Illustrated in Fig. 8.38(a) and similar to the discrete tuning technique described in Chapter 5 for LNAs, the idea is to place a bank of small capacitors, each having a value of C_u, in parallel with the tanks and switch them in or out to adjust the resonance frequency. We can also view V_{cont} as a "fine control" and the digital input to the capacitor bank as a "coarse control." Figure 8.38(b) shows the tuning behavior of the VCO as a function of both controls. The fine control provides a continuous but narrow range, whereas the coarse control shifts the continuous characteristic up or down.

The overall tuning range can be calculated as follows. With ideal switches and unit capacitors, the lowest frequency is obtained if all of the capacitors are switched in and the

---

5. The series resistance, R_S, decreases only slightly with ω because it is equal to the sum of the low-frequency component and the skin effect component, and because the latter varies with $\sqrt{\omega}$.

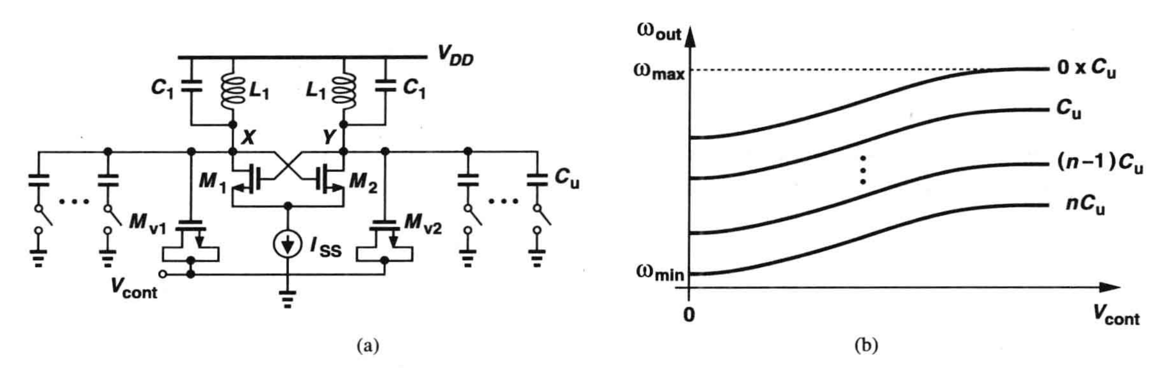


Figure 8.38 *(a) Discrete tuning by means of switched capacitors, (b) resulting characteristics.*

varactor is at its maximum value, C_{max}:

$$\omega_{min} = \frac{1}{\sqrt{L_1(C_1 + C_{max} + nC_u)}}. \tag{8.78}$$

The highest frequency occurs if the unit capacitors are switched out and the varactor is at its minimum value, C_{min}:

$$\omega_{max} = \frac{1}{\sqrt{L_1(C_1 + C_{min})}}. \tag{8.79}$$

Of course, as expressed by Eq. (8.77), the oscillation amplitude may vary considerably across this range, requiring "overdesign" at ω_{max} (or calibration) so as to obtain reasonable swings at ω_{min}.

Example 8.21

Consider the characteristics of Fig. 8.38(b) more carefully (Fig. 8.39). Does the continuous tuning range remain the same across the discrete tuning range? That is, can we say $\Delta\omega_{osc1} \approx \Delta\omega_{osc2}$?

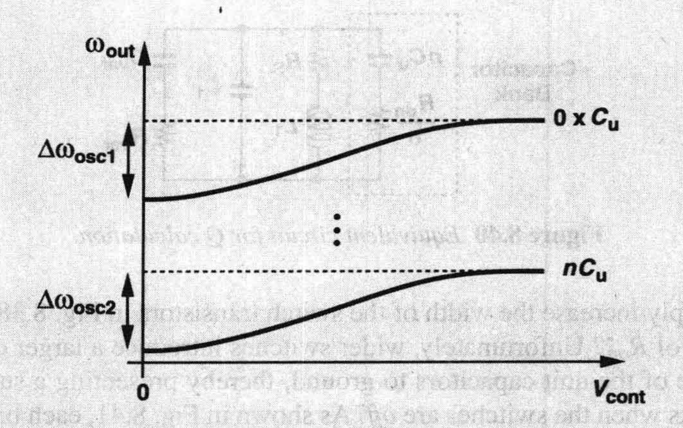


Figure 8.39 *Variation of fine tuning range.*

(Continues)

Example 8.21 *(Continued)*

Solution:

We expect $\Delta\omega_{osc1}$ to be greater than $\Delta\omega_{osc2}$ because, with nC_u switched into the tanks, the varactor sees a larger *constant* capacitance. In fact, from Eq. (8.53), we have

$$\Delta\omega_{osc1} \approx \frac{1}{\sqrt{L_1C_1}}\frac{C_{max}-C_{min}}{2C_1}, \tag{8.80}$$

and

$$\Delta\omega_{osc2} \approx \frac{1}{\sqrt{L_1(C_1+nC_u)}}\frac{C_{max}-C_{min}}{2(C_1+nC_u)}. \tag{8.81}$$

It follows that

$$\frac{\Delta\omega_{osc1}}{\Delta\omega_{osc2}} = \left(1+\frac{nC_u}{C_1}\right)^{3/2}. \tag{8.82}$$

This variation in K_{VCO} proves undesirable in PLL design.

The discrete tuning technique shown in Fig. 8.38(a) entails several difficult issues. First, the on-resistance, R_{on}, of the switches that control the unit capacitors degrades the Q of the tank. Applying Eq. (8.65) to the parallel combination shown in Fig. 8.40 and denoting $[(R_{on}/n)(nC_u)\omega]^{-1}$ by Q_{bank}, we have

$$\frac{1}{Q_{tot}} = \frac{1}{Q_L} + \frac{C_{var}}{C_1+C_{var}+nC_u}\frac{1}{Q_{var}} + \frac{nC_u}{C_1+C_{var}+nC_u}\frac{1}{Q_{bank}} \tag{8.83}$$

$$= \frac{R_S}{L_1\omega} + \frac{C_{var}}{C_1+C_{var}+nC_u}R_{var}C_{var}\omega + \frac{nC_u}{C_1+C_{var}+nC_u}R_{on}C_u\omega. \tag{8.84}$$

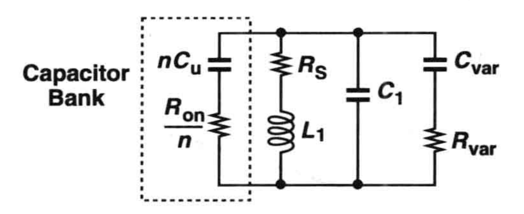


Figure 8.40 *Equivalent circuit for Q calculation.*

Can we simply increase the width of the switch transistors in Fig. 8.38(a) so as to minimize the effect of R_{on}? Unfortunately, wider switches introduce a larger capacitance from the bottom plate of the unit capacitors to ground, thereby presenting a substantial capacitance to the tanks when the switches are *off*. As shown in Fig. 8.41, each branch in the bank contributes a capacitance of $C_{GD}+C_{DB}$ to the tank if $C_u \gg C_{GD}+C_{DB}$. For n branches, therefore, C_1 incurs an additional constant component equal to $n(C_{GD}+C_{DB})$, further

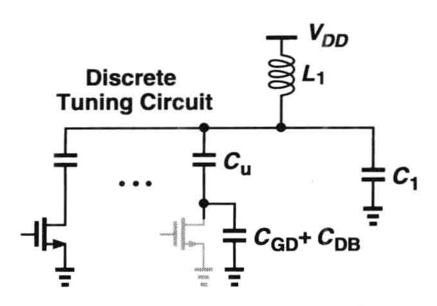


Figure 8.41 *Effect of switch parasitic capacitances.*

limiting the fine tuning range. For example, $\Delta\omega_{osc1}$ in Eq. (8.80) must be rewritten as

$$\Delta\omega_{osc1} \approx \frac{1}{\sqrt{L_1(C_1 + nC_{GD} + nC_{DB})}} \frac{C_{max} - C_{min}}{2(C_1 + nC_{GD} + nC_{DB})}. \tag{8.85}$$

This trade-off between the Q and the tuning range limits the use of discrete tuning.

The problem of switch on-resistance can be alleviated by exploiting the differential operation of the oscillator. Illustrated in Fig. 8.42(a), the idea is to place the main switch, S_1, between nodes A and B so that, with differential swings at these nodes, only *half* of R_{on1} appears in series with each unit capacitor [Fig. 8.42(b)]. This allows a twofold reduction in the switch width for a given resistance. Switches S_2 and S_3 are minimum-size devices, merely defining the CM level of A and B.

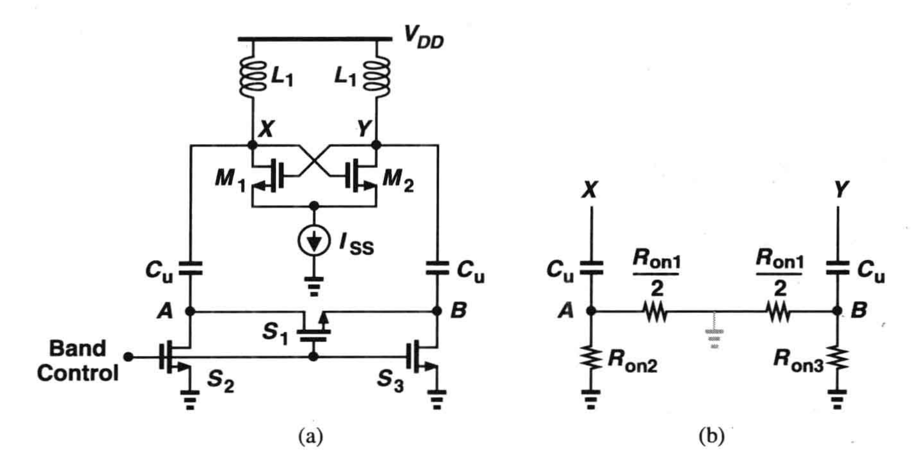


Figure 8.42 *(a) Use of floating switch, (b) equivalent circuit.*

The second issue in discrete tuning relates to potential "blind" zones. Suppose, as shown in Fig. 8.43(a), unit capacitor number j is switched out, creating a frequency change equal to $\omega_4 - \omega_2 \approx \omega_3 - \omega_1$, but the fine tuning range provided by the varactor, $\omega_4 - \omega_3$, is *less* than $\omega_4 - \omega_2$. Then, the oscillator fails to cover the range between ω_2 and ω_3 for any combination of fine and coarse controls.

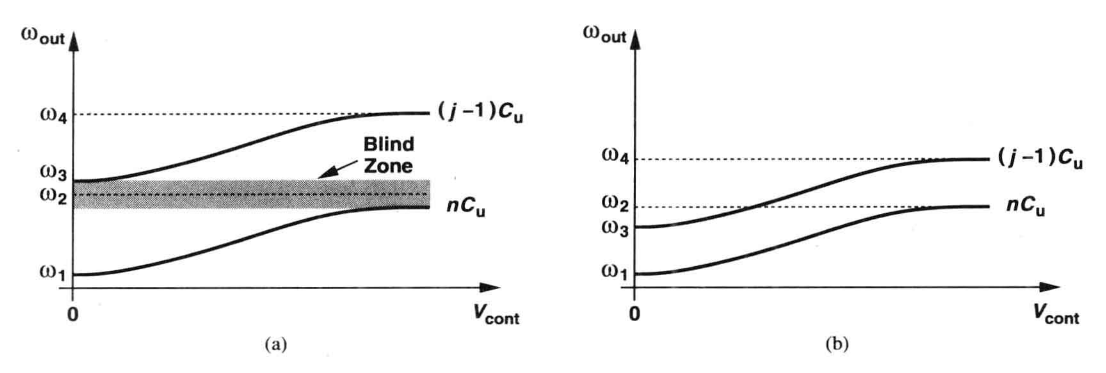


Figure 8.43 *(a) Blind zone in discrete tuning, (b) overlap between consecutive characteristics to avoid blind zone.*

To avoid blind zones, each two consecutive tuning characteristics must have some overlap. Depicted in Fig. 8.43(b), this precaution translates to smaller unit capacitors but a larger number of them and hence a complex layout. As explained in Chapter 13, the unit capacitors can be chosen unequal to mitigate this issue. Note that the overlap is also necessary to avoid excessive *variation* of K_{VCO} near the ends of each tuning curve. For example, around ω_2 in Fig. 8.43(b), the lower tuning curve flattens out, requiring that the upper one be used.

Some recent designs have employed oscillators with only discrete tuning. Called "digitally-controlled oscillators" (DCOs), such circuits must employ very fine frequency stops. Examples are described in [2].

In Chapter 13, we design and simulate a VCO with continuous and discrete tuning for 11a/g applications.

8.7 PHASE NOISE

The design of VCOs must deal with trade-offs among tuning range, phase noise, and power dissipation. Our study has thus far focused on the task of tuning. We now turn our attention to phase noise.

8.7.1 Basic Concepts

An ideal oscillator produces a perfectly-periodic output of the form $x(t) = A\cos\omega_c t$. The zero crossings occur at exact integer multiples of $T_c = 2\pi/\omega_c$. In reality, however, the noise of the oscillator devices randomly perturbs the zero crossings. To model this perturbation, we write $x(t) = A\cos[\omega_c t + \phi_n(t)]$, where $\phi_n(t)$ is a small random phase quantity that deviates the zero crossings from integer multiples of T_c. Figure 8.44 illustrates the two waveforms in the time domain. The term $\phi_n(t)$ is called the "phase noise."

The waveforms in Fig. 8.44 can also be viewed from another, slightly different, perspective. We can say that the *period* remains constant if $x(t) = A\cos\omega_c t$ but varies randomly if $x(t) = A\cos[\omega_c t + \phi_n(t)]$ (as indicated by $T_1, \ldots, T_m$ in Fig. 8.44). In other

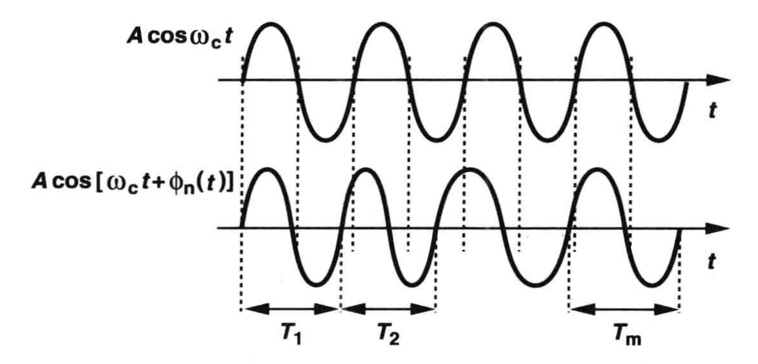


Figure 8.44 *Output waveforms of an ideal and a noisy oscillator.*

words, the *frequency* of the waveform is constant in the former case but varies randomly in the latter. This observation leads to the spectrum of the oscillator output. For $x(t) = A\cos\omega_c t$, the spectrum consists of a single impulse at ω_c [Fig. 8.45(a)], whereas for $x(t) = A\cos[\omega_c t + \phi_n(t)]$ the frequency experiences random variations, i.e., it departs from ω_c occasionally. As a consequence, the impulse is "broadened" to represent this random departure [Fig. 8.45(b)].

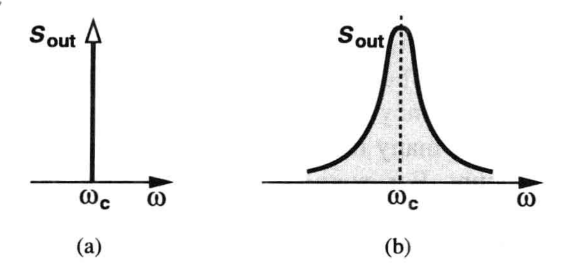


Figure 8.45 *Output spectra of (a) an ideal, and (b) a noisy oscillator.*

Example 8.22

Explain why the broadened impulse cannot assume the shape shown in Fig. 8.46.

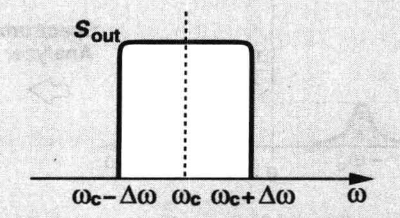


Figure 8.46 *Flat spectrum around oscillation frequency.*

(Continues)

Example 8.22 *(Continued)*

Solution:

This spectrum occurs if the oscillator frequency has *equal* probability of appearing anywhere between $\omega_c - \Delta\omega$ and $\omega_c + \Delta\omega$. However, we intuitively expect that the oscillator prefers ω_c to other frequencies, thus spending lesser time at frequencies that are farther from ω_c. This explains the declining phase noise "skirts" in Fig. 8.45(b).

Our focus on noise in the zero crossings rather than noise on the amplitude arises from the assumption that the latter is removed by hard switching in stages following the oscillator. For example, the switching transistors in an active mixer spend little time near equilibrium, "masking" most of the LO amplitude noise for the rest of the time.

The spectrum of Fig. 8.45(b) can be related to the time-domain expression. Since $\phi_n(t) \ll 1$ rad,

$$x(t) = A\cos[\omega_c t + \phi_n(t)] \tag{8.86}$$

$$\approx A\cos\omega_c t - A\sin\omega_c t \sin[\phi_n(t)] \tag{8.87}$$

$$\approx A\cos\omega_c t - A\phi_n(t)\sin\omega_c t. \tag{8.88}$$

That is, the spectrum of $x(t)$ consists of an impulse at ω_c and the spectrum of $\phi_n(t)$ *translated* to a center frequency of ω_c. Thus, the declining skirts in Fig. 8.45(b) in fact represent the behavior of $\phi_n(t)$ in the frequency domain.

In phase noise calculations, many factors of 2 or 4 appear at different stages and must be carefully taken into account. For example, as illustrated in Fig. 8.47, (1) since $\phi_n(t)$ in Eq. (8.88) is multiplied by $\sin\omega_c t$, its *power* spectral density, $S_{\phi n}$, is multiplied by 1/4 as it is translated to $\pm\omega_c$; (2) A spectrum analyzer measuring the resulting spectrum *folds* the negative-frequency spectrum atop the positive-frequency spectrum, raising the spectral density by a factor of 2.

How is the phase noise quantified? Since the phase noise falls at frequencies farther from ω_c, it must be specified at a certain "frequency offset," i.e., a certain difference with

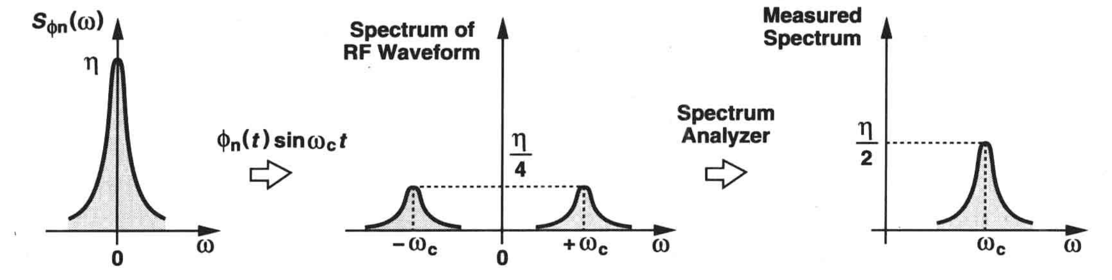


Figure 8.47 *Various factors of 4 and 2 that arise in conversion of noise to phase noise.*

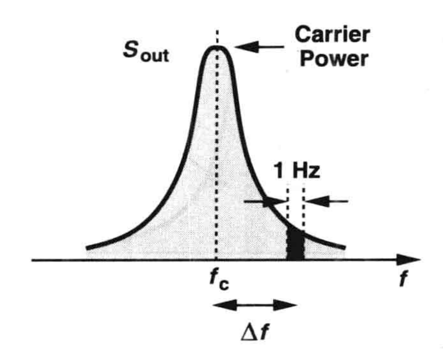


Figure 8.48 *Specification of phase noise.*

respect to ω_C. As shown in Fig. 8.48, we consider a 1-Hz bandwidth of the spectrum at an offset of Δf, measure the power in this bandwidth, and normalize the result to the "carrier power." The carrier power can be viewed as the peak of the spectrum or (more rigorously) as the power given by Eq. (8.86), namely, $A^2/2$. For example, the phase noise of an oscillator in GSM applications must fall below -115 dBc/Hz at 600-kHz offset. Called "dB with respect to the carrier," the unit dBc signifies normalization of the noise power to the carrier power.

Example 8.23

At high carrier frequencies, it is difficult to measure the noise power in a 1-Hz bandwidth. Suppose a spectrum analyzer measures a noise power of -70 dBm in a 1-kHz bandwidth at 1-MHz offset. How much is the phase noise at this offset if the average oscillator output power is -2 dBm?

Solution:

Since a 1-kHz bandwidth carries $10\log(1000\text{ Hz}) = 30$ dB higher noise than a 1-Hz bandwidth, we conclude that the noise power in 1 Hz is equal to -100 dBm. Normalized to the carrier power, this value translates to a phase noise of -98 dBc/Hz.

In practice, the phase noise reaches a constant floor at large frequency offsets (beyond a few megahertz) (Fig. 8.49). We call the regions near and far from the carrier the "close-in" and the "far-out" phase noise, respectively, although the border between the two is vague.

8.7.2 Effect of Phase Noise

To understand the effect of phase noise in RF systems, let us consider the receiver front end shown in Fig. 8.50(a) and study the downconverted spectrum. Referring to the ideal

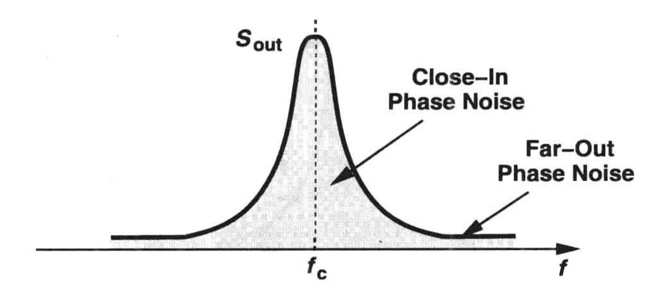


Figure 8.49 *Close-in and far-out phase noise.*

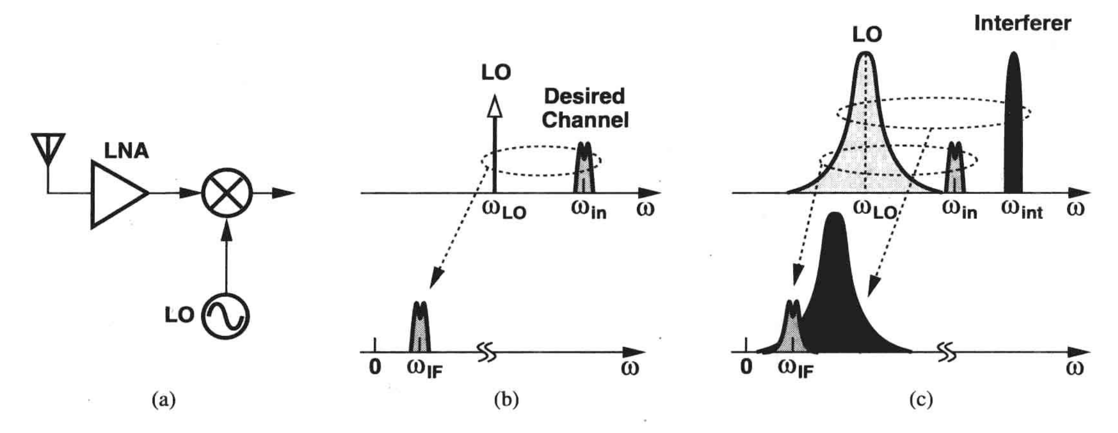


Figure 8.50 *(a) Receive front end, (b) downconversion with an ideal LO, (c) downconversion with a noisy LO (reciprocal mixing).*

case depicted in Fig. 8.50(b), we observe that the desired channel is convolved with the impulse at ω_{LO}, yielding an IF signal at $\omega_{IF} = \omega_{in} - \omega_{LO}$. Now, suppose the LO suffers from phase noise and the desired signal is accompanied by a large interferer. As illustrated in Fig. 8.50(c), the convolution of the desired signal and the interferer with the noisy LO spectrum results in a *broadened* downconverted interferer whose noise skirt corrupts the desired IF signal. This phenomenon is called "reciprocal mixing."

Reciprocal mixing becomes critical in receivers that may sense large interferers. The LO phase noise must then be so small that, when integrated across the desired channel, it produces negligible corruption.

Example 8.24

A GSM receiver must withstand an interferer located three channels away from the desired channel and 45 dB higher. Estimate the maximum tolerable phase noise of the LO if the corruption due to reciprocal mixing must remain 15 dB below the desired signal.

Example 8.24 (*Continued*)

Solution:

Figure 8.51 depicts the downconverted spectrum. The total noise power introduced by the interferer in the desired channel is equal to

$$P_{n,tot} = \int_{f_L}^{f_H} S_n(f)df, \tag{8.89}$$

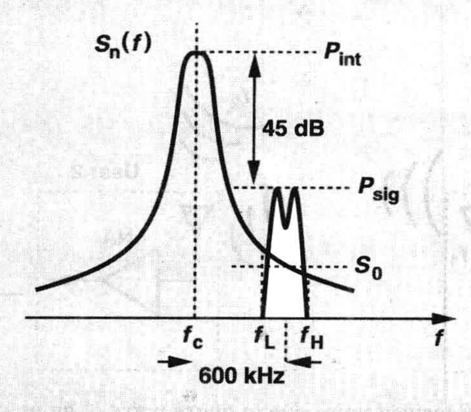


Figure 8.51 *Example of reciprocal mixing.*

where $S_n(f)$ denotes the broadened spectrum of the interferer and f_L and f_H are the lower and upper ends of the desired channel, respectively. For simplicity, we assume $S_n(f)$ is relatively flat in this bandwidth and equal to S_0, obtaining $P_{n,tot} = S_0(f_H - f_L)$. Thus,

$$\text{SNR} = \frac{P_{sig}}{S_0(f_H - f_L)}, \tag{8.90}$$

which must be at least 15 dB. In other words,

$$10 \log \frac{S_0}{P_{sig}} = -15 \text{ dB} - 10 \log(f_H - f_L). \tag{8.91}$$

Since the interferer is convolved with the LO phase noise (S_0), it must be normalized to P_{int}. Noting that $10 \log(P_{int}/P_{sig}) = 45$ dB, we rewrite Eq. (8.91) as

$$10 \log \frac{S_0}{P_{int}} = -15 \text{ dB} - 10 \log(f_H - f_L) - 45 \text{ dB}. \tag{8.92}$$

If $f_H - f_L = 200$ kHz, then

$$10 \log \frac{S_0}{P_{int}} = -113 \text{ dBc/Hz at } 600\text{–kHz offset}. \tag{8.93}$$

(*Continues*)

Example 8.24 (*Continued*)

In practice, the phase noise skirt is not constant from f_L to f_H, calling for a more accurate calculation. We perform a more accurate analysis in Chapter 13.

Phase noise also manifests itself in transmitters. Shown in Fig. 8.52 is a scenario where two users are located in close proximity, with user #1 transmitting a high-power signal at f_1 and user #2 receiving this signal and a weak signal at f_2. If f_1 and f_2 are only a few channels apart, the phase noise skirt masking the signal received by user #2 greatly corrupts it even *before* downconversion.

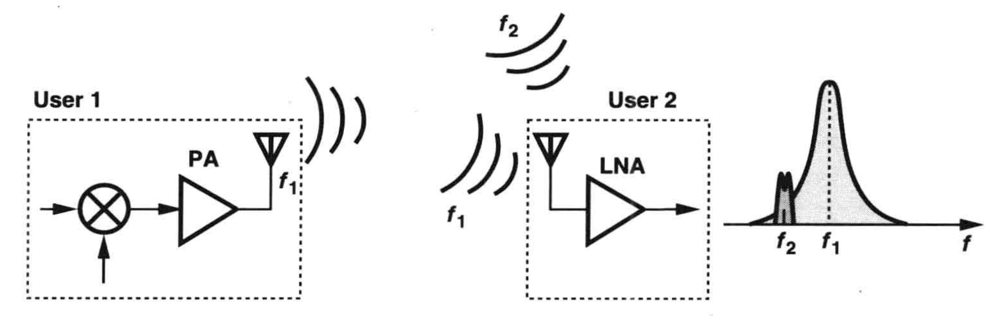


Figure 8.52 *Received noise due to phase noise of an unwanted signal.*

Example 8.25

A student reasons that, if the interferer at f_1 in Fig. 8.52 is so large that its phase noise corrupts the reception by user #2, then it also heavily *compresses* the receiver of user #2. Is this true?

Solution:

Not necessarily. As evidenced by Example 8.24, an interferer, say, 50 dB above the desired signal produces phase noise skirts that are not negligible. For example, the desired signal may have a level of −90 dBm and the interferer, −40 dBm. Since most receivers' 1-dB compression point is well above −40 dBm, user #2's receiver experiences no desensitization, but the phenomenon in Fig. 8.52 is still critical.

The LO phase noise also corrupts phase-modulated signals in the process of upconversion or downconversion. Since the phase noise is indistinguishable from phase (or frequency) modulation, the mixing of the signal with a noisy LO in the TX or RX path corrupts the information carried by the signal. For example, a QPSK signal containing phase noise can be expressed as

$$x_{QPSK}(t) = A\cos\left[\omega_c t + (2k+1)\frac{\pi}{4} + \phi_n(t)\right] \quad k = 0, \cdots, 3 \tag{8.94}$$

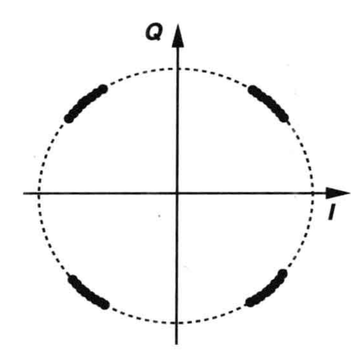


Figure 8.53 *Corruption of a QPSK signal due to phase noise.*

revealing that the amplitude is unaffected by phase noise. Thus, the constellation points experience only random rotation around the origin (Fig. 8.53). If large enough, phase noise and other nonidealities move a constellation point to another quadrant, creating an error.

Example 8.26

Which points in a 16-QAM constellation are most sensitive to phase noise?

Solution:

Consider the four points in the top right quadrant (Fig. 8.54). Points B and C can tolerate a rotation of 45° before they move to adjacent quadrants. Points A and D, on the other hand, can rotate by only $\theta = \tan^{-1}(1/3) = 18.4^\circ$. Thus, the eight outer points near the I and Q axes are most sensitive to phase noise.

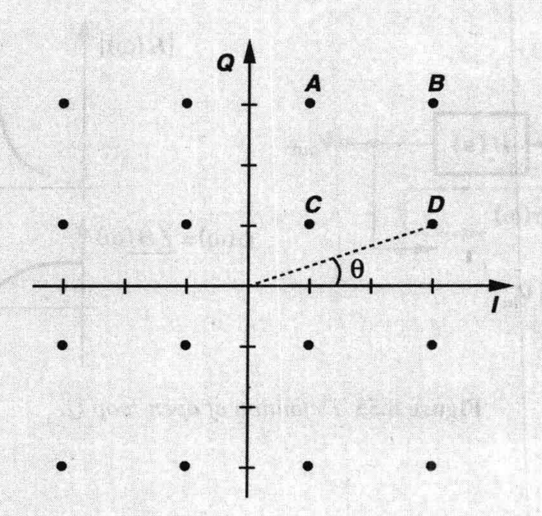


Figure 8.54 *16-QAM constellation for study of effect of phase noise.*

8.7.3 Analysis of Phase Noise: Approach I

Oscillator phase noise has been under study for decades [3]–[17], leading to a multitude of analysis techniques in the frequency and time domains. The calculation of phase noise by hand still remains tedious, but simulation tools such as Cadence's SpectreRF have greatly simplified the task. Nonetheless, a solid understanding of the mechanisms that give rise to phase noise proves essential to oscillator design. In this section, we analyze these mechanisms. In particular, we must answer two important questions: (1) how much and at what point in an oscillation cycle does each device "inject" noise? (2) how does the injected noise produce phase noise in the output voltage waveform?

Q of an Oscillator In Chapters 2 and 7, we derived various expressions for the Q of an LC tank. We know intuitively that a high Q signifies a sharper resonance, i.e., a higher selectivity. Another definition of the Q that is especially well-suited to oscillators is illustrated in Fig. 8.55. Here, the circuit is viewed as a feedback system and the *phase* of the *open-loop* transfer function, $\phi(\omega)$, is examined at the resonance frequency, ω_0. The "open-loop" Q is defined as

$$Q = \frac{\omega_0}{2}\left|\frac{d\phi}{d\omega}\right|. \tag{8.95}$$

This definition offers an interesting insight if we recall that for steady oscillation, the total phase shift around the loop must be 360° (or zero). Suppose the noise injected by the devices attempts to deviate the frequency from ω_0. From Fig. 8.55, such a deviation translates to a change in the total phase shift around the loop, violating Barkhausen's criterion and forcing the oscillator to return to ω_0. The larger the slope of $\phi(j\omega)$, the greater is this "restoration" force; i.e., oscillators with a high open-loop Q tend to spend less time at frequencies other than ω_0. In Problem 8.10, we prove that this definition of Q coincides with our original definition, $Q = R_p/(L\omega)$, for a CS stage loaded by a second-order tank.

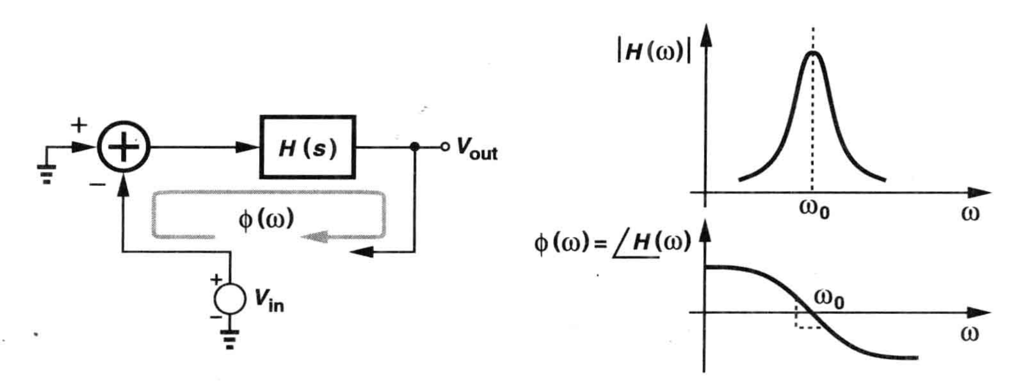


Figure 8.55 *Definition of open-loop Q.*

Example 8.27

Compute the open-loop Q of a cross-coupled LC oscillator.

Solution:

We construct the open-loop circuit as shown in Fig. 8.56 and note that $V_{out}/V_X = V_X/V_{in}$ and hence $H(s) = V_{out}/V_{in} = (V_X/V_{in})^2$. Thus, the phase of V_{out}/V_{in} is equal to twice the phase of V_X/V_{in}. Since at $s = j\omega$,

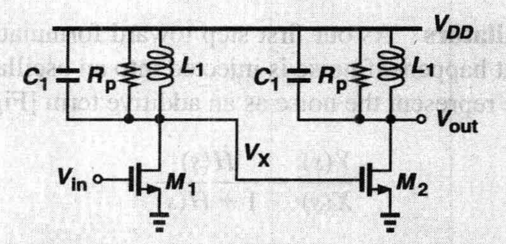


Figure 8.56 *Open-loop model of a cross-coupled oscillator.*

$$\frac{V_X}{V_{in}}(j\omega) = \frac{-jg_m R_p L_1 \omega}{R_p(1 - L_1 C_1 \omega^2) + jL_1\omega}, \tag{8.96}$$

we have

$$\angle H(j\omega) = 2\left[-\frac{\pi}{2} - \tan^{-1}\frac{L_1\omega}{R_p(1 - L_1 C_1 \omega^2)}\right]. \tag{8.97}$$

Differentiating both sides with respect to ω, calculating the result at $\omega_0 = (\sqrt{L_1 C_1})^{-1}$, and multiplying it by $\omega_0/2$, we obtain

$$\left|\frac{\omega_0}{2}\frac{d\angle H(j\omega)}{d\omega}\right|_{\omega 0} = 2R_p C_1 \omega_0 \tag{8.98}$$

$$= 2Q_{tank}, \tag{8.99}$$

where Q_{tank} denotes the Q of each tank. This result is to be expected: the cascade of frequency-selective stages makes the phase transition sharper than that of one stage.

While the open-loop Q indicates how much an oscillator "rejects" the noise, the phase noise depends on three other factors as well: (1) the *amount* of noise that different devices inject, (2) the point in time during a cycle at which the devices inject noise (some parts of the waveform are more sensitive than others), and (3) the output voltage swing (carrier power). We elaborate on these as we analyze phase noise.

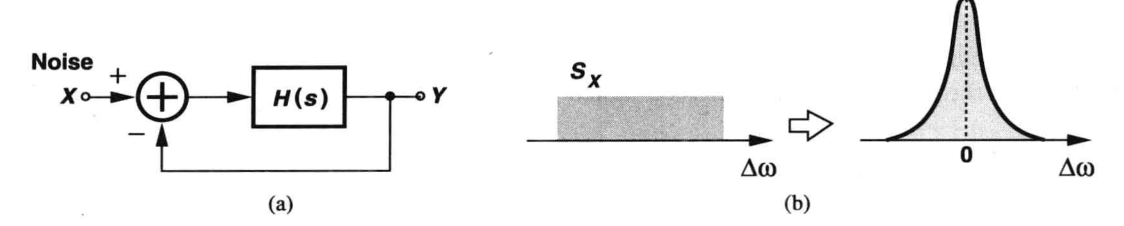


Figure 8.57 *(a) Oscillator model, (b) noise shaping in oscillator.*

Noise Shaping in Oscillators As our first step toward formulating the phase noise, we wish to understand what happens if noise is injected into an oscillatory circuit. Employing the feedback model, we represent the noise as an additive term [Fig. 8.57(a)] and write

$$\frac{Y(s)}{X(s)} = \frac{H(s)}{1 + H(s)}. \tag{8.100}$$

In the vicinity of the oscillation frequency, i.e., at $\omega = \omega_0 + \Delta\omega$, we can approximate $H(j\omega)$ with the first two terms in its Taylor series:

$$H(j\omega) \approx H(j\omega_0) + \Delta\omega \frac{dH}{d\omega}. \tag{8.101}$$

If $H(j\omega_0) = -1$ and $\Delta\omega dH/d\omega \ll 1$, then Eq. (8.100) reduces to

$$\frac{Y}{X}(j\omega_0 + j\Delta\omega) \approx \frac{-1}{\Delta\omega \dfrac{dH}{d\omega}}. \tag{8.102}$$

In other words, as shown in Fig. 8.57(b), the noise spectrum is "shaped" by

$$\left|\frac{Y}{X}(j\omega_0 + j\Delta\omega)\right|^2 = \frac{1}{\Delta\omega^2 |\dfrac{dH}{d\omega}|^2}. \tag{8.103}$$

To determine the shape of $|dH/d\omega|^2$, we write $H(j\omega)$ in polar form, $H(j\omega) = |H| \exp(j\phi)$ and differentiate with respect to ω,

$$\frac{dH}{d\omega} = \left(\frac{d|H|}{d\omega} + j|H|\frac{d\phi}{d\omega}\right) \exp(j\phi). \tag{8.104}$$

It follows that

$$\left|\frac{dH}{d\omega}\right|^2 = \left|\frac{d|H|}{d\omega}\right|^2 + \left|\frac{d\phi}{d\omega}\right|^2 |H|^2. \tag{8.105}$$

This equation leads to a general definition of Q [4], but we limit our study here to simple LC oscillators. Note that (a) in an LC oscillator, the term $|d|H|/d\omega|^2$ is much less than

$|d\phi/d\omega|^2$ in the vicinity of the resonance frequency, and (b) $|H|$ is close to unity for steady oscillations. The right-hand side of Eq. (8.105) therefore reduces to $|d\phi/d\omega|^2$, yielding

$$\left|\frac{Y}{X}(j\omega_0 + j\Delta\omega)\right|^2 = \frac{1}{\dfrac{\omega_0^2}{4}\left|\dfrac{d\phi}{d\omega}\right|^2}\frac{\omega_0^2}{4\Delta\omega^2}. \tag{8.106}$$

From (8.95),

$$\left|\frac{Y}{X}(j\omega_0 + j\Delta\omega)\right|^2 = \frac{1}{4Q^2}\left(\frac{\omega_0}{\Delta\omega}\right)^2. \tag{8.107}$$

Known as "Leeson's Equation" [3], this result reaffirms our intuition that the open-loop Q signifies how much the oscillator rejects the noise.

Example 8.28

A student designs the cross-coupled oscillator of Fig. 8.58 with $2/g_m = 2R_p$, reasoning that the tank now has infinite Q and hence the oscillator produces no phase noise![6] Explain the flaw in this argument. (This circuit is similar to that in Fig. 8.21(b), but with the tank components renamed.)

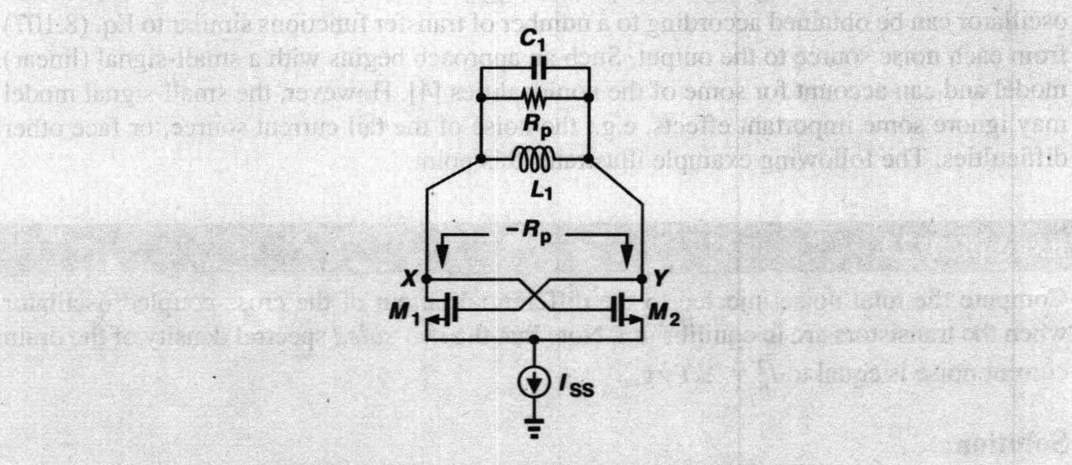


Figure 8.58 *Apparently infinite Q in an oscillator.*

Solution:

The Q in Eq. (8.107) is the *open-loop* Q, i.e., $\omega_0/2$ times the slope of the phase of the *open-loop* transfer function, which was calculated in Example 8.27. The "closed-loop" Q does not carry much meaning.

6. The center tap of L_1 is tied to V_{DD} but not shown.

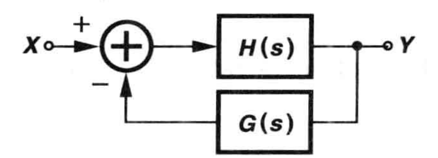


Figure 8.59 *Noise shaping in a general oscillator.*

In Problem 8.11, we prove that, if the feedback path has a transfer function $G(s)$ (Fig. 8.59), then

$$\left|\frac{Y}{X}(j\omega_0 + j\Delta\omega)\right|^2 = \frac{1}{4Q^2}\left(\frac{\omega_0}{\Delta\omega}\right)^2\left|\frac{1}{G(j\omega_0)}\right|^2, \tag{8.108}$$

where the open-loop Q is given by

$$Q = \frac{\omega_0}{2}\left|\frac{d(GH)}{d\omega}\right|. \tag{8.109}$$

Linear Model The foregoing development suggests that the total noise at the output of an oscillator can be obtained according to a number of transfer functions similar to Eq. (8.107) from each noise source to the output. Such an approach begins with a small-signal (linear) model and can account for some of the nonidealities [4]. However, the small-signal model may ignore some important effects, e.g., the noise of the tail current source, or face other difficulties. The following example illustrates this point.

Example 8.29

Compute the total noise injected to the differential output of the cross-coupled oscillator when the transistors are in equilibrium. Note that the *two-sided* spectral density of the drain current noise is equal to $\overline{I_n^2} = 2kT\gamma g_m$.

Solution:

Let us first determine the Norton equivalent of the cross-coupled pair. From Fig. 8.60(a), the reader can show that the short-circuit output current, I_X, is equal to half of the noise current of each transistor: $I_X = (I_{n2} - I_{n1})/2$. Thus, as shown in Fig. 8.60(b), the output noise is obtained as

$$\overline{V_{n,out}^2} = \left(\overline{I_X^2} + \frac{2kT}{R_p}\right)\frac{R^2L_1^2\omega^2}{R^2(1 - L_1C_1\omega^2)^2 + L_1^2\omega^2}, \tag{8.110}$$

Example 8.29 (*Continued*)

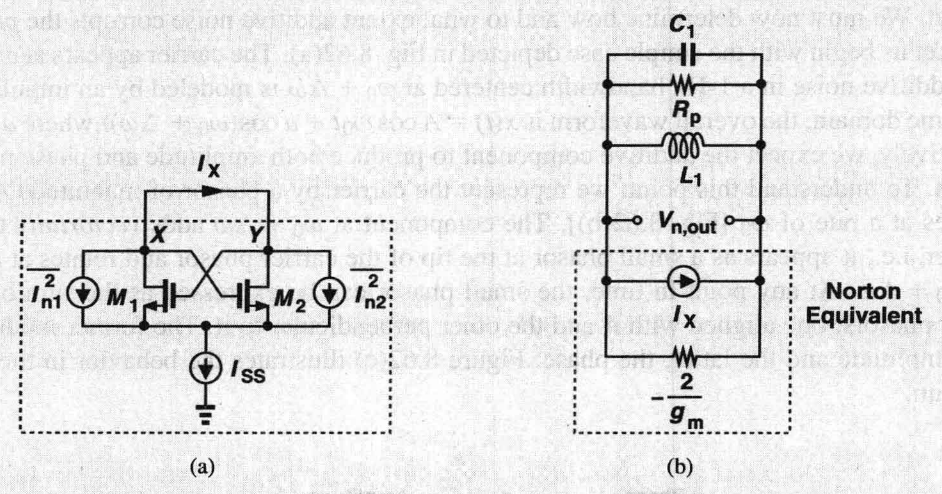


Figure 8.60 *(a) Circuit for finding Norton noise equivalent of cross-coupled pair, (b) overall model of oscillator.*

where $R = (-2/g_m)||R_p$. Since I_{n1} and I_{n2} are uncorrelated, $\overline{I_X^2} = \left(\overline{I_{n1}^2} + \overline{I_{n2}^2}\right)/4 = kT\gamma g_m$ and hence

$$\overline{V_{n,out}^2} = \left(kT\gamma g_m + \frac{2kT}{R_p}\right) \frac{R^2 L_1^2 \omega^2}{R^2(1 - L_1 C_1 \omega^2)^2 + L_1^2 \omega^2}. \tag{8.111}$$

Figure 8.61 plots the spectrum of $\overline{V_{n,out}^2}$. Unfortunately, this result contradicts Leeson's equation. As explained in Section 8.3, g_m is typically quite higher than $2/R_p$ and hence $R \neq \infty$. Thus, as $\omega \to \omega_0$, $\overline{V_{n,out}^2}$ does not go to infinity. This is another difficulty arising from the linear model.

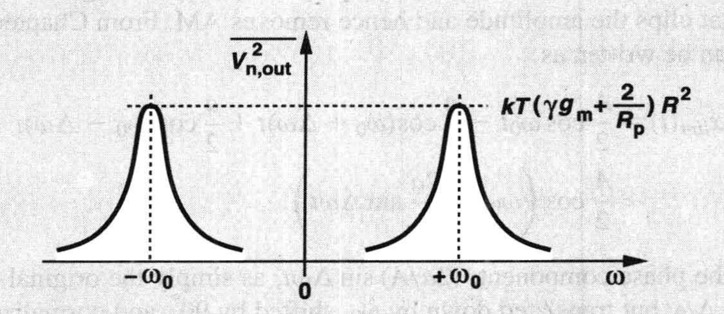


Figure 8.61 *Output spectrum due to transistor noise currents.*

Conversion of Additive Noise to Phase Noise The result expressed by (8.107) and exemplified by (8.111) yields the total noise that is *added* to the oscillation waveform at the output. We must now determine how and to what extent additive noise corrupts the *phase*.

Let us begin with the simple case depicted in Fig. 8.62(a). The carrier appears at ω_0 and the additive noise in a 1-Hz bandwidth centered at $\omega_0 + \Delta\omega$ is modeled by an impulse. In the time domain, the overall waveform is $x(t) = A\cos\omega_0 t + a\cos(\omega_0 + \Delta\omega)t$ where $a \ll A$. Intuitively, we expect the additive component to produce both amplitude and phase modulation. To understand this point, we represent the carrier by a phasor of magnitude A that rotates at a rate of ω_0 [Fig. 8.62(b)]. The component at $\omega_0 + \Delta\omega$ adds *vectorially* to the carrier, i.e., it appears as a small phasor at the tip of the carrier phasor and rotates at a rate of $\omega_0 + \Delta\omega$. At any point in time, the small phasor can be expressed as the sum of two other phasors, one aligned with A and the other perpendicular to it. The former modulates the amplitude and the latter, the phase. Figure 8.62(c) illustrates the behavior in the time domain.

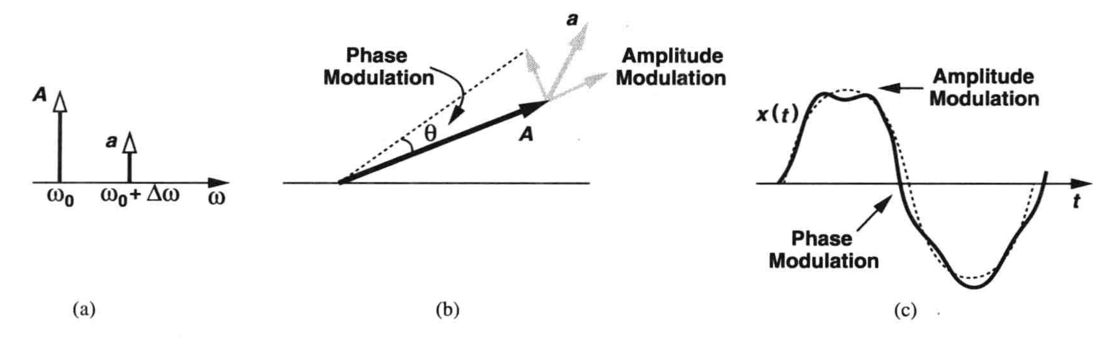


Figure 8.62 *(a) Addition of a small sideband to a sinusoid, (b) phasor diagram showing both AM and PM, (c) time-domain waveform.*

In order to compute the phase modulation resulting from a small sinusoid at $\omega_0 + \Delta\omega$, we make two important observations. First, as described in Chapter 3, the spectrum of Fig. 8.62(a) can be written as the sum of an AM signal and an FM signal. Second, the phase of the overall signal is obtained by applying the composite signal to a hard limiter, i.e., a circuit that clips the amplitude and hence removes AM. From Chapter 3, the output of the limiter can be written as

$$x_{lim}(t) = \frac{A}{2}\cos\omega_0 t - \frac{a}{2}\cos(\omega_0 + \Delta\omega)t + \frac{a}{2}\cos(\omega_0 - \Delta\omega)t \tag{8.112}$$

$$\approx \frac{A}{2}\cos\left(\omega_0 t - \frac{2a}{A}\sin\Delta\omega t\right). \tag{8.113}$$

We recognize the phase component, $(2a/A)\sin\Delta\omega t$, as simply the original additive component at $\omega_0 + \Delta\omega$, but translated down by ω_0, shifted by 90°, and normalized to $A/2$. We therefore expect that narrowband random additive noise in the vicinity of ω_0 results in a phase whose spectrum has the *same* shape as that of the additive noise but translated by ω_0 and normalized to $A/2$.

This conjecture can be proved analytically. We write $x(t) = A\cos\omega_0 t + n(t)$, where $n(t)$ denotes the narrowband additive noise (voltage or current). It can be proved that

narrowband noise in the vicinity of ω_0 can be expressed in terms of its *quadrature components* [9]:

$$n(t) = n_I(t)\cos\omega_0 t - n_Q(t)\sin\omega_0 t, \tag{8.114}$$

where $n_I(t)$ and $n_Q(t)$ have the same spectrum, which [for real $n(t)$] is equal to the spectrum of $n(t)$ but translated down by ω_0 (Fig. 8.63) and doubled in spectral density. It follows that

$$x(t) = [A + n_I(t)]\cos\omega_0 t - n_Q(t)\sin\omega_0 t. \tag{8.115}$$

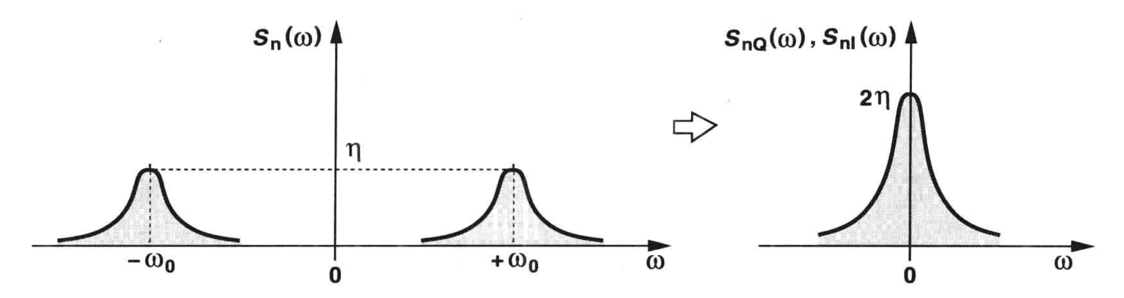


Figure 8.63 *Narrowband noise and spectrum of its quadrature components.*

Expressing Eq. (8.115) in polar form, we have

$$x(t) = \sqrt{[A + n_I(t)]^2 + n_Q^2(t)}\cos\left[\omega_0 t + \tan^{-1}\frac{n_Q(t)}{A + n_I(t)}\right]. \tag{8.116}$$

Since $n_I(t), n_Q(t) \ll A$, the phase component is equal to

$$\phi_n(t) \approx \frac{n_Q(t)}{A}, \tag{8.117}$$

as postulated previously. It follows that

$$S_{\phi n}(\omega) = \frac{S_{nQ}(\omega)}{A^2}. \tag{8.118}$$

Note that A is the peak (not the rms) amplitude of the carrier. In Problem 8.12, we prove that half of the noise power is carried by the AM sidebands and the other half by the PM sidebands.

We are ultimately interested in the spectrum of the RF waveform, $x(t)$, but excluding its AM noise. We have

$$x(t) \approx A\cos\left[\omega_0 t + \frac{n_Q(t)}{A}\right] \tag{8.119}$$

$$\approx A\cos\omega_0 t - n_Q(t)\sin\omega_0 t. \tag{8.120}$$

Thus, the power spectral density of $x(t)$ consists of two impulses at $\pm\omega_0$, each with a power of $A^2/4$, and $S_{nQ}/4$ centered around $\pm\omega_0$. As shown in Fig. 8.64, a spectrum analyzer folds

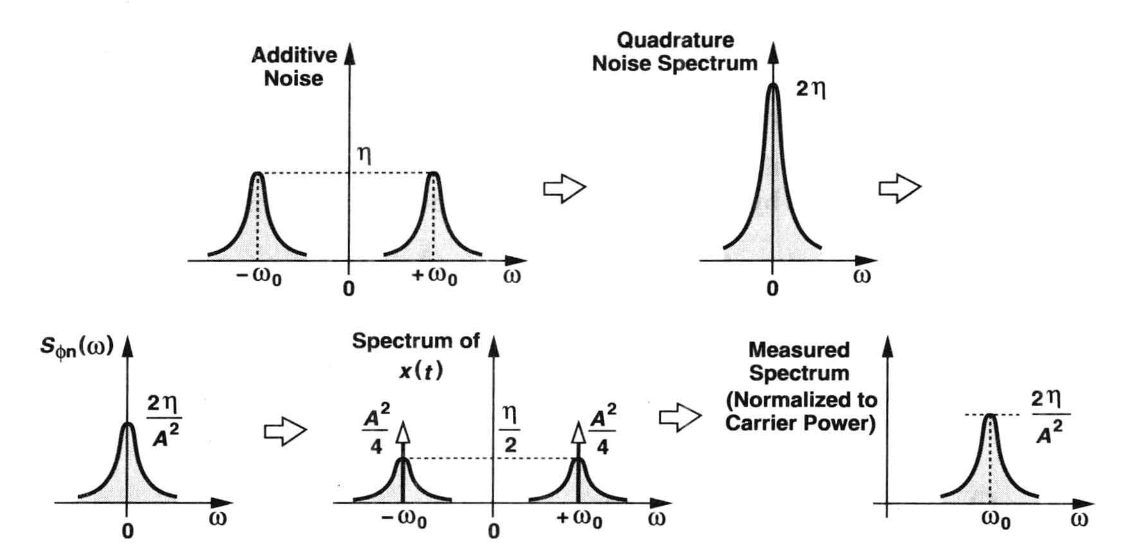


Figure 8.64 *Summary of conversion of additive noise to phase noise.*

the negative- and positive-frequency contents. After the folding, we normalize the phase noise to the total carrier power, $A^2/2$.

The foregoing development can be summarized as follows (Fig. 8.64). Additive noise around $\pm\omega_0$ having a two-sided spectral density with a peak of η results in a phase noise spectrum around ω_0 having a normalized one-sided spectral density with a peak of $2\eta/A^2$, where A is the *peak* amplitude of the carrier.

Cyclostationary Noise The derivations leading to Eq. (8.111) have assumed that the noise of each transistor can be represented by a *constant* spectral density; however, as the transistors experience large-signal excursions, their transconductance and hence noise power varies. Since oscillators perform this noise modulation periodically, we say such noise sources are "cyclostationary," i.e., their spectrum varies periodically. We begin our analysis with an observation made in Chapter 6 regarding cyclostationary white noise: white noise multiplied by a periodic envelope in the time domain remains white. For example, if white noise is switched on and off with 50% duty cycle, the result is still white but has half the spectral density.

In order to study the effect of cyclostationary noise, we return to the original cross-coupled oscillator and, from Fig. 8.65(a), recognize that (1) when V_X reaches a maximum and V_Y a minimum, M_1 turns off, injecting no noise; (2) when M_1 and M_2 are near equilibrium, they inject maximum noise current, with a total two-sided spectral density of $kT\gamma g_m$, where g_m is the equilibrium transconductance; (3) when V_X reaches a minimum and V_Y a maximum, M_1 is on but degenerated by the tail current (while M_2 is off), injecting little noise to the output. We therefore conclude that the total noise current experiences an envelope having *twice* the oscillation frequency and swinging between zero and unity [Fig. 8.65(b)].

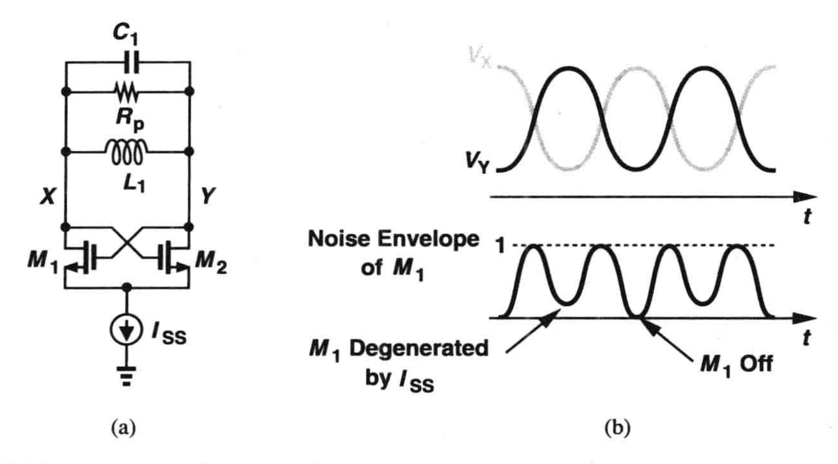


Figure 8.65 *(a) General cross-coupled oscillator, (b) envelope of transistor noise waveforms.*

The noise envelope waveform can be determined by simulations, but let us approximate the envelope by a sinusoid, $0.5 \cos 2\omega_0 t + 0.5$. The reader can show that white noise multiplied by such an envelope results in white noise with three-eighths the spectral density. Thus, we simply multiply the noise contribution of M_1 and M_2, $kT\gamma g_m$, by 3/8.

How about the noise of the tanks? We observe that the noise of R_p in Fig. 8.58 is stationary. In other words, the two-sided tank noise contribution is equal to $2kT/R_p$ (but only half of this value is converted to phase noise).

Time-Varying Resistance In addition to cyclostationary noise, the time variation of the resistance presented by the cross-coupled pair also complicates the analysis. However, since we have taken a "macroscopic" view of cyclostationary noise and modeled it by an equivalent white noise, we may consider a *time average* of the resistance as well.

We have noted that the resistance seen between the drains of M_1 and M_2 in Fig. 8.65(a) periodically varies from $-2/g_m$ to nearly infinity. The corresponding conductance, G, thus swings between $-g_m/2$ and nearly zero (Fig. 8.66), exhibiting a certain average, $-G_{avg}$. The value of $-G_{avg}$ is readily obtained as the first term of the Fourier expansion of the conductance waveform.

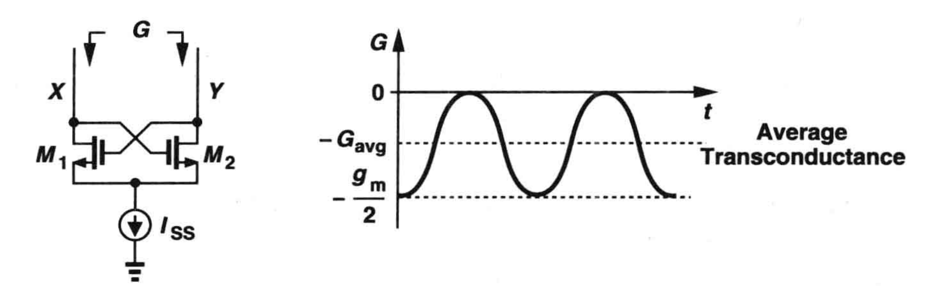


Figure 8.66 *Time variation of conductance of cross-coupled pair.*

What can we say about $-G_{avg}$? If $-G_{avg}$ is not sufficient to compensate for the loss of the tank, R_p, then the oscillation decays. Conversely, if $-G_{avg}$ is more than enough, then the oscillation amplitude grows. In the steady state, therefore, $G_{avg} = 1/R_p$. This is a powerful observation: regardless of the transistor dimensions and the value of the tail current, G_{avg} must remain equal to R_p.

Example 8.30

What happens to the conductance waveform and G_{avg} if the tail current is increased?

Solution:

Since G_{avg} must remain equal to $1/R_p$, the waveform changes shape such that it has greater excursions but still the same average value. As shown in Fig. 8.67(a), a larger tail current leads to a greater peak transconductance, $-g_{m2}/2$, while increasing the time that the transconductance spends near zero so that the average is constant. That is, the transistors are at equilibrium for a shorter amount of time [Fig. 8.67(b)].

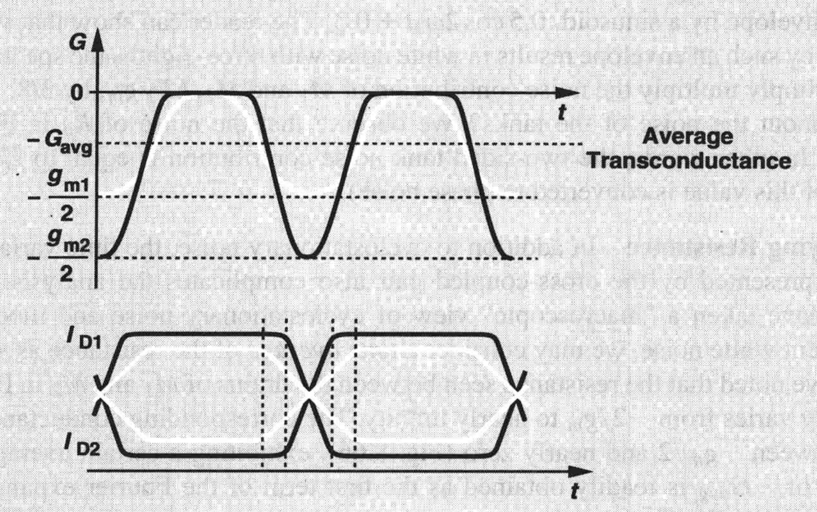


Figure 8.67 *Effect of increasing tail current on (a) transconductance, and (b) oscillation waveforms.*

Phase Noise Computation We now consolidate our formulations of (a) conversion of additive noise to phase noise, (b) cyclostationary noise, and (c) time-varying resistance. Our analysis proceeds as follows:

1. We compute the average spectral density of the noise current injected by the cross-coupled pair. If a sinusoidal envelope is assumed, the two-sided spectral density amounts to $kT\gamma g_m \times (3/8)$, where g_m denotes the equilibrium transconductance of each transistor.
2. To this we add the noise current of R_p, obtaining $(3/8)kT\gamma g_m + 2kT/R_p$.

3. We multiply the above spectral density by the squared magnitude of the net impedance seen between the output nodes. Since $G_{avg} = 1/R_p$, the average resistance is *infinite*, leaving only L_1 and C_1 in Fig. 8.65(a). That is,

$$\overline{V_{n,out}^2} = kT\left(\frac{3}{8}\gamma g_m + \frac{2}{R_p}\right)\frac{L_1^2\omega^2}{(1 - L_1C_1\omega^2)^2}, \tag{8.121}$$

which, for $\omega = \omega_0 + \Delta\omega$ and $\Delta\omega \ll \omega_0$, reduces to

$$\overline{V_{n,out}^2} = kT\left(\frac{3}{8}\gamma g_m + \frac{2}{R_p}\right)\frac{1}{4C_1^2\Delta\omega^2}. \tag{8.122}$$

The factor of 3/8 depends on the noise envelope waveform and must be obtained by careful simulations.

4. From Fig. 8.64, we divide this result by $A^2/2$ to obtain the one-sided phase noise spectrum around ω_0. Note that in Fig. 8.65(a), $A = (4/\pi)(I_{SS}R_p/2) = (2/\pi)I_{SS}/R_p$ and $R_p = QL_1\omega_0$.[7] It follows that

$$S(\Delta\omega) = \frac{\pi^2}{2}\frac{kT}{I_{SS}^2}\left(\frac{3}{8}\gamma g_m + \frac{2}{R_p}\right)\frac{\omega_0^2}{4Q^2\Delta\omega^2}. \tag{8.123}$$

As the tail current and hence the output swings increase, I_{SS}^2 rises much more sharply than g_m, thereby lowering the phase noise (so long as the transistors do not enter the deep triode region).

A closer examination of the cross-coupled oscillator reveals that the phase noise is in fact independent of the transconductance of the transistors [10, 11, 17]. This can be qualitatively justified as follows. Suppose the widths of the two transistors are increased while the output voltage swing and frequency are kept constant. The transistors can now steer their tail current with a smaller voltage swing, thus experiencing sharper current switching (Fig. 8.68). That is, M_1 and M_2 spend less time injecting noise into the tank. However, the higher transconductance translates to a higher amount of injected noise, as evident from the noise envelope. It turns out that the decrease in the width and the increase in the height of the noise envelope pulses cancel each other, and g_m can be simply replaced with $2/R_p$ in the above equation [10, 11, 17]:

$$S(\Delta\omega) = \frac{\pi^2}{R_p}\frac{kT}{I_{SS}^2}\left(\frac{3}{8}\gamma + 1\right)\frac{\omega_0^2}{4Q^2\Delta\omega^2}. \tag{8.124}$$

Problem of Tail Capacitance What happens if one of the transistors enters the deep triode region? As depicted in Fig. 8.69(a), the corresponding tank is temporarily connected to the tail capacitance through the on-resistance of the transistor, degrading the Q.

7. The differential resistance, R_p, can be viewed as two resistors of value $R_p/2$ tied to V_{DD}. The peak single-ended swing is therefore equal to $(2/\pi)(R_p/2)I_{SS}$.

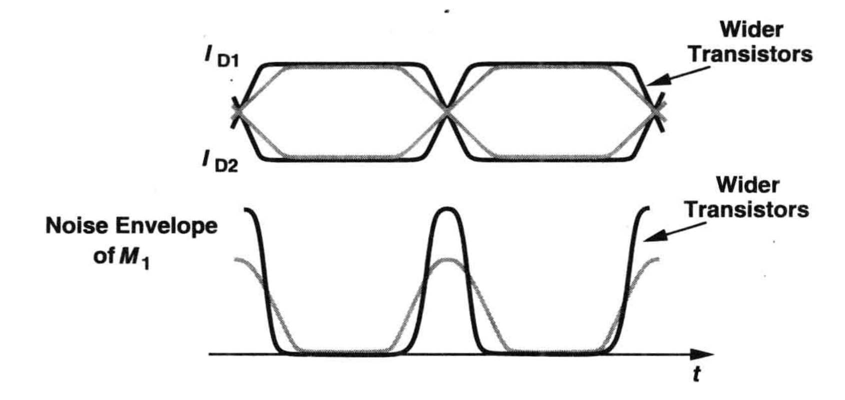


Figure 8.68 *Oscillator waveforms for different transistor widths.*

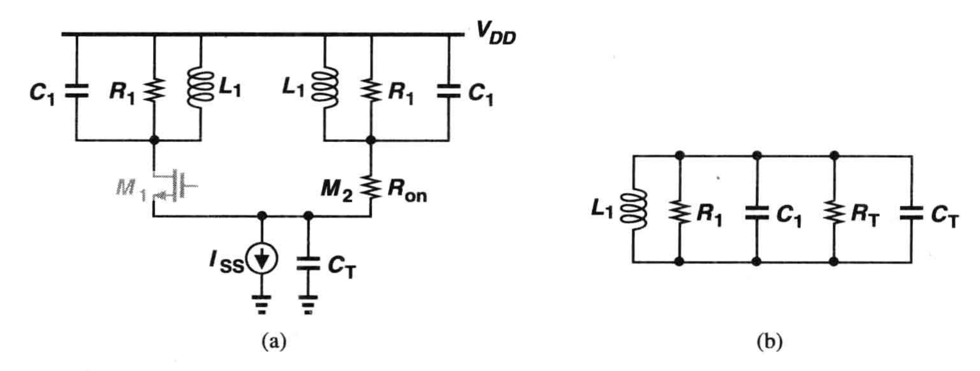


Figure 8.69 *(a) Oscillator with one transistor in deep triode region, (b) equivalent circuit of the tank.*

Transforming the series combination of R_{on} and C_T to a parallel circuit, we obtain the equivalent network shown in Fig. 8.69(b), where $R_T = (R_{on}C_T^2\omega_0^2)^{-1}$. If R_T is comparable to R_1 and each transistor remains in the deep triode region for an appreciable fraction of the period, then the Q degrades significantly. Equivalently, the noise injected by M_2 rises considerably [17].

The key result here is that, as the tail current is increased, the (relative) phase noise continues to decline up to the point where the transistors enter the triode region. Beyond this point, a higher tail current raises the output swing more gradually, but the overall tank Q begins to fall, yielding no significant improvement in the phase noise. Of course, this trend depends on the value of C_T and may be pronounced only in some designs. This capacitance may be large due to the parasitics of I_{SS} or it may be added deliberately to shunt the noise of I_{SS} to ground (Section 8.7.5).

Example 8.31

How does the above phenomenon manifest itself in the top-biased topology of Fig. 8.30(a)?

Solution:

In this case, each transistor entering the deep triode region provides a direct resistive path to ground. Since node P is also at (ac) ground, the tank Q heavily deteriorates. We thus expect this topology to suffer more severely if M_1 or M_2 enters the deep triode region.

8.7.4 Analysis of Phase Noise: Approach II

The approach described in this section follows that in [6]. Consider an ideal LC tank that, due to an initial condition, produces a sinusoidal output [Fig. 8.70(a)]. During the oscillation, L_1 and C_1 exchange the initial energy, with the former carrying the entire energy at the zero crossings and the latter, at the peaks. Let us assume that the circuit begins with an initial voltage of V_0 across the capacitor. Now, suppose an impulse of current is injected into the oscillating tank at the peak of the output voltage [Fig. 8.70(b)], producing a voltage step across C_1. If[8]

$$I_{in}(t) = I_1\delta(t - t_1), \tag{8.125}$$

then the additional energy gives rise to a larger oscillation amplitude:

$$V_p = V_0 + \frac{I_1}{C_1}. \tag{8.126}$$

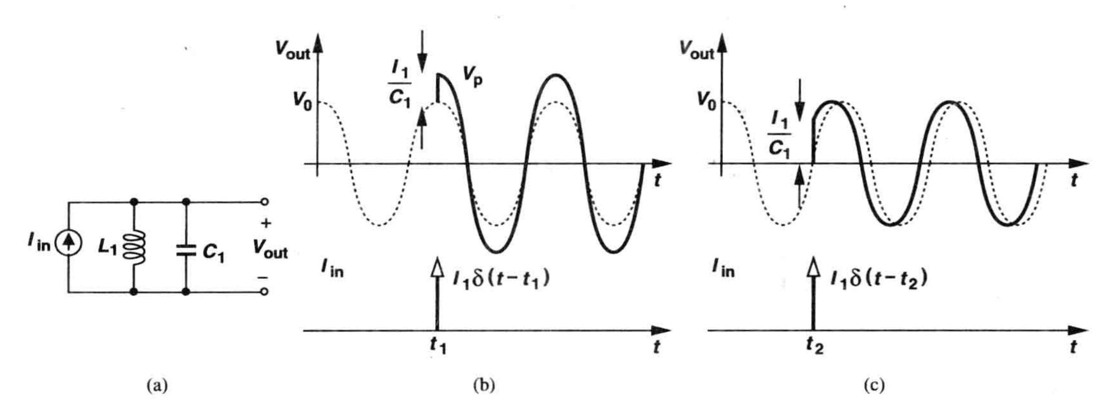


Figure 8.70 *(a) Ideal tank with a current impulse, (b) effect of impulse injection at peak of waveform, (c) effect of impulse injection at zero crossing of waveform.*

8. Note that I_1 in this equation is in fact a charge quantity because it denotes the area under the impulses.

The key point here is that the injection at the peak does not disturb the *phase* of the oscillation (as shown in the example below).

Next, let us assume the impulse of current is injected at a zero crossing point. A voltage step is again created but leading to a *phase* jump [Fig. 8.70(c)]. Since the voltage jumps from 0 to I_1/C_1, the phase is disturbed by an amount equal to $\sin^{-1}[I_1/(C_1V_0)]$. We therefore conclude that noise creates only amplitude modulation if injected at the peaks and only phase modulation if injected at the zero crossings.

Example 8.32

Explain how the effect of the current impulse can be determined analytically.

Solution:

The linearity of the tank allows the use of superposition for the injected currents (the inputs) and the voltage waveforms (the outputs). The output waveform consists of two sinusoidal components, one due to the initial condition[9] (the oscillation waveform) and another due to the impulse. Figure 8.71 illustrates these components for two cases: if injected at t_1, the impulse leads to a sinusoid exactly in phase with the original component, and if injected at t_2, the impulse produces a sinusoid 90° out of phase with respect to the original component. In the former case, the peaks are unaffected, and in the latter, the zero crossings.

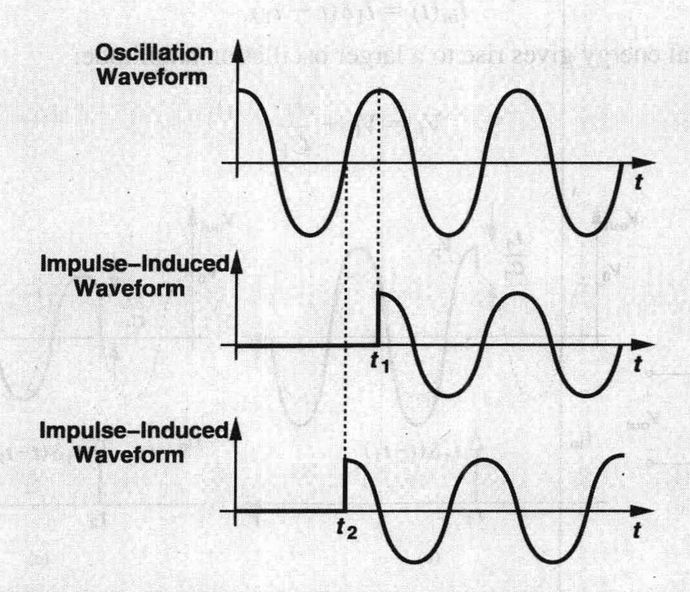


Figure 8.71 *Computation of impulse response using superposition.*

The foregoing observations suggest the need for a method of quantifying how and when each source of noise in an oscillator "hits" the output waveform. While the transistors turn on and off, a noise source may only appear near the peaks of the output voltage, contributing

9. The initial condition in the tank can also be created by a current impulse and hence does not make the system nonlinear.

negligible phase noise, whereas another may hit the zero crossings, producing substantial phase noise. To this end, we define a linear, *time-variant* system from each noise source to the output phase. The linearity property is justified because the noise levels are very small, and the time variance is necessary to capture the effect of the time at which the noise appears at the output.

For a linear, time-variant system, the convolution property holds, but the impulse response varies with time. Thus, the output phase in response to a noise $n(t)$ is given by

$$\phi(t) = h(t, \tau) * n(t), \tag{8.127}$$

where $h(t, \tau)$ is the time-variant impulse response from $n(t)$ to $\phi(t)$. In an oscillator, $h(t, \tau)$ varies *periodically*: as illustrated in Fig. 8.72, a noise impulse injected at $t = t_1$ or integer multiples of the period thereafter produces the same phase change. Now, the task of output phase noise calculation consists of computing $h(t, \tau)$ for each noise source and convolving it with the noise waveform. The impulse response, $h(t, \tau)$, is called the "impulse sensitivity function" (ISF) in [6].

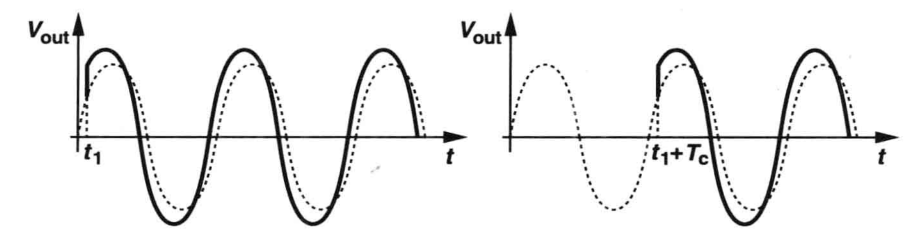


Figure 8.72 *Periodic impulse response in an oscillator.*

Example 8.33

Explain how the LC tank of Fig. 8.70(a) has a time-variant behavior even though the inductor and the capacitor values remain constant.

Solution:

The time variance arises from the finite *initial condition* (e.g., the initial voltage across C_1). With a zero initial condition, the circuit begins with a zero output, exhibiting a time-invariant response to the input.

Example 8.34

Compute the phase impulse response for the lossless LC tank of Fig. 8.70(a).

Solution:

We invoke the superposition perspective of Example 8.32 and wish to calculate the phase change resulting from a current impulse at an arbitrary time t_1 (Fig. 8.73). The overall output voltage can be expressed as

$$V_{out}(t) = V_0 \cos \omega_0 t + \Delta V[\cos \omega_0 (t - t_1)]u(t - t_1), \tag{8.128}$$

(Continues)

Example 8.34 *(Continued)*

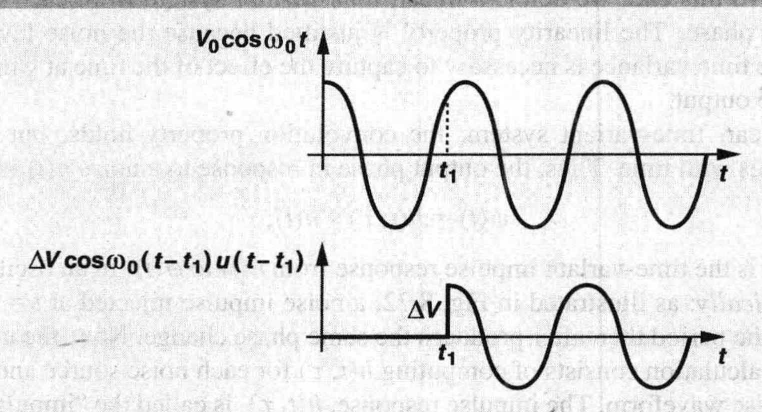


Figure 8.73 *Waveforms for computation of phase impulse response of a tank.*

where ΔV is given by the area under the impulse (I_1 in Fig. 8.70) divided by C_1. For $t \geq t_1$, V_{out} is equal to the sum of two sinusoids:

$$V_{out}(t) = V_0 \cos\omega_0 t + \Delta V \cos\omega_0(t - t_1) \quad t \geq t_1 \tag{8.129}$$

which, upon expansion of the second term and regrouping, reduces to

$$V_{out}(t) = (V_0 + \Delta V \cos\omega_0 t_1)\cos\omega_0 t + \Delta V \sin\omega_0 t_1 \sin\omega_0 t \quad t \geq t_1. \tag{8.130}$$

The phase of the output is therefore equal to

$$\phi_{out} = \tan^{-1}\frac{\Delta V \sin\omega_0 t_1}{V_0 + \Delta V \cos\omega_0 t_1} \quad t \geq t_1. \tag{8.131}$$

Interestingly, ϕ_{out} is *not* a linear function of ΔV in general. But, if $\Delta V \ll V_0$, then

$$\phi_{out} \approx \frac{\Delta V}{V_0}\sin\omega_0 t_1 \quad t \geq t_1. \tag{8.132}$$

If normalized to the area under the input impulse (I_1), this result yields the impulse response:

$$h(t, t_1) = \frac{1}{C_1 V_0}\sin\omega_0 t_1 u(t - t_1). \tag{8.133}$$

As expected, $h(t, t_1)$ is zero at $t_1 = 0$ (at the peak of $V_0 \cos\omega_0 t$) and maximum at $t_1 = \pi/(2\omega_0)$ (at the zero crossing of $V_0 \cos\omega_0 t$).

Let us now return to Eq. (8.127) and determine how the convolution is carried out. It is instructive to begin with a linear, time-invariant system. A given input, $x(t)$, can be approximated by a series of time-domain impulses, each carrying the energy of $x(t)$ in a

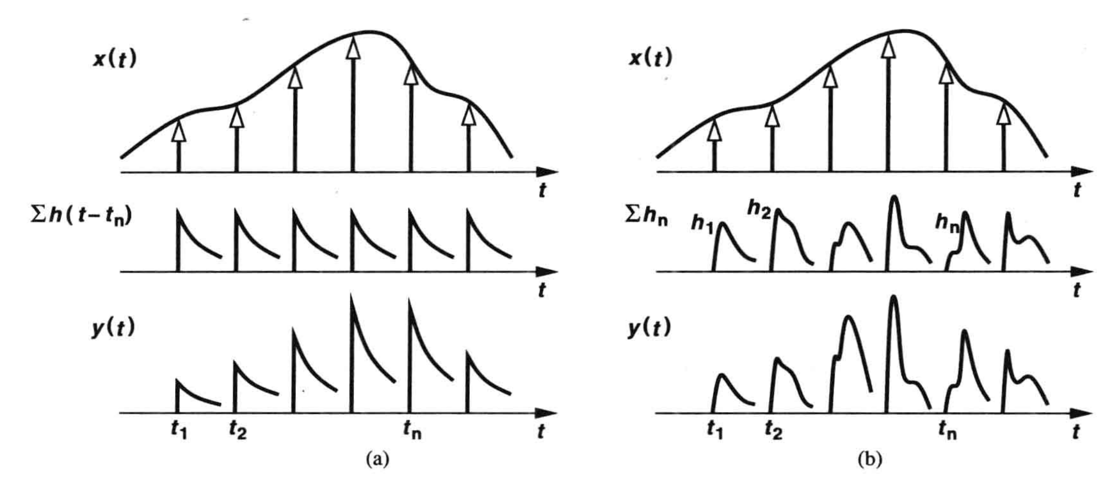


Figure 8.74 *Convolution in a (a) time-invariant, and (b) time-variant linear system.*

short time span [Fig. 8.74(a)]:

$$x(t) \approx \sum_{n=-\infty}^{+\infty} x(t_n)\delta(t - t_n). \tag{8.134}$$

Each impulse produces the time-invariant impulse response of the system at t_n. Thus, $y(t)$ consists of time-shifted replicas of $h(t)$, each scaled in amplitude according to the corresponding value of $x(t)$:

$$y(t) \approx \sum_{n=-\infty}^{+\infty} x(t_n)h(t - t_n) \tag{8.135}$$

$$= \int_{-\infty}^{+\infty} x(\tau)h(t - \tau)d\tau. \tag{8.136}$$

Now, consider the time-variant system shown in Fig. 8.74(b). In this case, the time-shifted versions of $h(t)$ may be different, and we denote them by $h_1(t)$, $h_2(t)$, ..., $h_n(t)$, with the understanding that $h_j(t)$ is the impulse response in the vicinity of t_j. It follows that

$$y(t) \approx \sum_{n=-\infty}^{+\infty} x(t_n)h_n(t). \tag{8.137}$$

How do we express these impulse responses as a continuous-time function? We simply write them as $h(t, \tau)$, where τ is the specific time shift. For example, $h_1(t) = h(t, 1\text{ ns})$, $h_2(t) = h(t, 2\text{ ns})$, etc. Thus,

$$y(t) = \int_{-\infty}^{+\infty} x(\tau)h(t, \tau)d\tau. \tag{8.138}$$

Example 8.35

Determine the phase noise resulting from a current, $i_n(t)$, having a white spectrum, $S_i(f)$, that is injected into the tank of Fig. 8.70(a).

Solution:

From Eqs. (8.133) and (8.138),

$$\phi_n(t) = \int_{-\infty}^{+\infty} i_n(\tau)\frac{1}{C_1 V_0}\sin\omega_0\tau u(t-\tau)d\tau \tag{8.139}$$

$$= \frac{1}{C_1 V_0}\int_{-\infty}^{t} i_n(\tau)\sin\omega_0\tau d\tau. \tag{8.140}$$

If $i_n(t)$ is white, so is $g(t) = i_n(t)\sin\omega_0 t$ (why?), but with *half* the spectral density of $i_n(t)$:

$$S_g(f) = \frac{1}{2}S_i(f). \tag{8.141}$$

Our task therefore reduces to finding the transfer function of the system shown in Fig. 8.75(a). To this end, we note that (1) the impulse response of this system is simply equal to $(C_1V_0)^{-1}u(t)$, and (2) the Fourier transform of $u(t)$ is given by $(j\omega)^{-1} + \pi\delta(\omega)$. We ignore $\pi\delta(\omega)$ as it contains energy at only $\omega = 0$ and write

$$S_{\phi n}(f) = |H(j\omega)|^2 S_g(f) \tag{8.142}$$

$$= \frac{1}{C_1^2V_0^2}\frac{1}{(2\pi f)^2}\frac{S_i(f)}{2}. \tag{8.143}$$

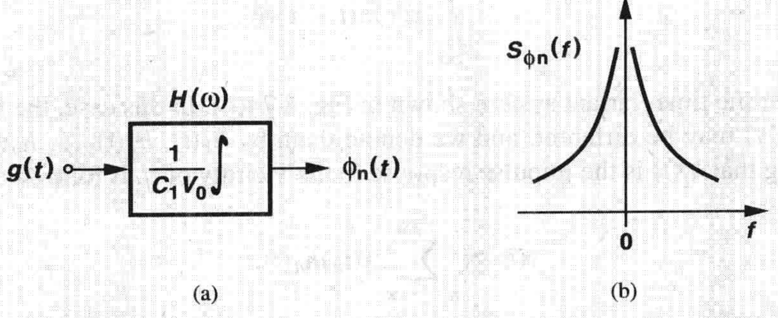


Figure 8.75 *(a) Equivalent system for conversion of $g(t)$ to $\phi_n(t)$, (b) resulting phase noise spectrum.*

As expected, the relative phase noise is inversely proportional to the oscillation peak amplitude, V_0. Depicted in Fig. 8.75(b), this equation agrees with the noise-shaping concept described in Section 8.7.3: the spectrum of $A\cos[\omega_0 t + \phi_n(t)]$ contains $S_{\phi n}(f)$ but shifted to a center frequency of $\pm\omega_0$. Figure 8.76 summarizes the mechanisms that convert the

Example 8.35 *(Continued)*

injected noise current to phase noise. For clarity, the white noise near $\pm\omega_0$ is shown as three narrowband segments.

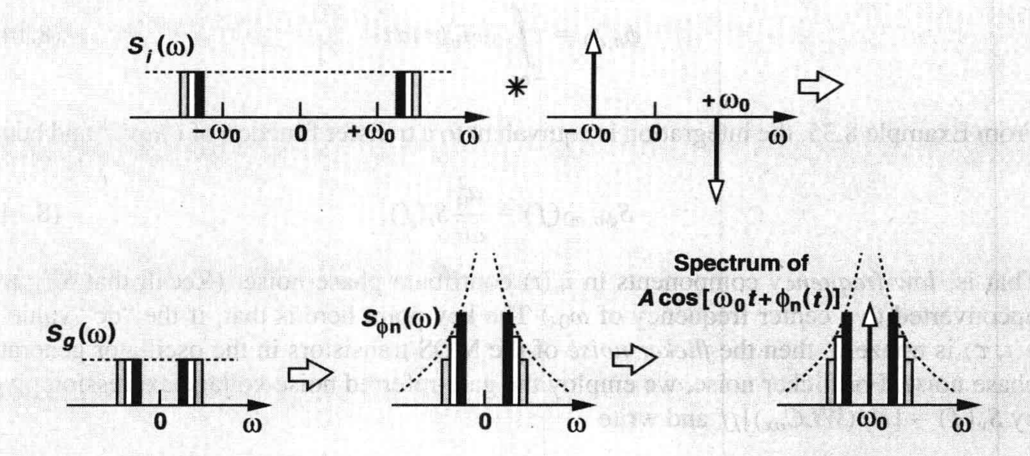


Figure 8.76 *Summary of conversion of injected noise to phase noise around the carrier.*

The reader may find the foregoing example confusing: if the lossless tank with its nonzero initial condition is viewed as an oscillator with infinite Q, why is the phase noise not *zero*? This confusion is resolved if we recognize that, as $Q \to \infty$, the width and bias current of the transistor needed to sustain oscillation become infinitesimally small. The transistor thus injects nearly zero noise; i.e., if $i_n(t)$ represents transistor noise, $S_i(f)$ approaches zero.

Example 8.36

Which frequency components in $i_n(t)$ in the above example contribute significant phase noise?

Solution:

Since $i_n(t)$ is multiplied by $\sin\omega_0 t$, noise components around ω_0 are translated to the vicinity of zero frequency and subsequently appear in Eq. (8.143) (Fig. 8.76). Thus, for a sinusoidal phase impulse response (ISF), only noise frequencies near ω_0 contribute significant phase noise. The reader is encouraged to repeat the translations in Fig. 8.76 for a low-frequency component in $S_i(\omega)$.

Effect of Flicker Noise Due to its periodic nature, the impulse response of oscillators can be expressed as a Fourier series:

$$h(t,\tau) = [a_0 + a_1\cos(\omega_0 t + \phi_1) + a_2\cos(2\omega_0 t + \phi_2) + \cdots]u(t-\tau), \tag{8.144}$$

where a_0 is the average (or "dc") value of $h(t, \tau)$. In the LC tank studied above, $a_j = 0$ for all $j \neq 1$, but in general this may not be true. In particular, suppose $a_0 \neq 0$. Then, the corresponding phase noise in response to an injected noise $i_n(t)$ is equal to

$$\phi_{n,a0} = \int_{-\infty}^{t} a_0 i_n(\tau) d\tau. \tag{8.145}$$

From Example 8.35, the integration is equivalent to a transfer function of $(j\omega)^{-1}$ and hence

$$S_{\phi n,a0}(f) = \frac{a_0^2}{\omega^2} S_i(f). \tag{8.146}$$

That is, *low-frequency* components in $i_n(t)$ contribute phase noise. (Recall that $S_{\phi n,a0}$ is upconverted to a center frequency of ω_0.) The key point here is that, if the "dc" value of $h(t, \tau)$ is nonzero, then the *flicker noise* of the MOS transistors in the oscillator generates phase noise. For flicker noise, we employ the gate-referred noise voltage expression given by $S_v(f) = [K/(WLC_{ox})]/f$ and write

$$S_{\phi n,a0}(f) = \frac{a_0^2}{4\pi^2} \frac{K}{WLC_{ox}} \frac{1}{f^3}. \tag{8.147}$$

Note that in this case, a_0 represents the dc term of the impulse response from the *gate voltage* of the transistors to the output phase. Since a_0 relates to the symmetry of $h(t, \tau)$, low upconversion of $1/f$ noise requires a circuit design that exhibits an odd-symmetric $h(t, \tau)$ [6]. However, the $1/f$ noise of different transistors in the circuit may see different impulse responses, and it may therefore be impossible to minimize the upconversion of *all* $1/f$ noise sources. For example, in the circuit of Fig. 8.35, it is possible to make $h(t, \tau)$ from the gates of M_1-M_4 to the output symmetric, but not from the tail current source to the output. As I_{SS} slowly fluctuates, so does the output CM level and hence the oscillation frequency. In general, the phase noise spectrum assumes the shape shown in Fig. 8.77.

Noise around Higher Harmonics Let us now turn our attention to the remaining terms in Eq. (8.144). As mentioned in Example 8.35, $a_1 \cos(\omega_0 t + \phi_1)$ translates noise frequencies around ω_0 to the vicinity of zero and into phase noise. By the same token, $a_m \cos(m\omega_0 t + \phi_j)$ converts noise components around $m\omega_0$ in $i_n(t)$ to phase noise. Figure 8.78 illustrates this behavior [6].

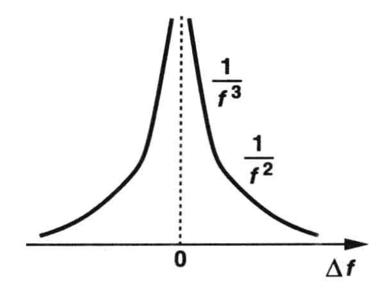


Figure 8.77 *Phase noise profile showing regions arising from flicker and white noise.*

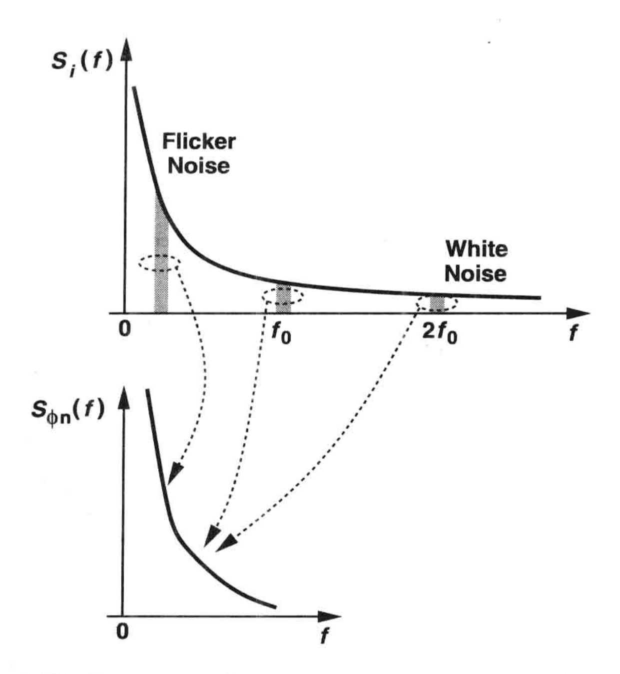


Figure 8.78 *Conversion of various noise components to phase noise.*

Cyclostationary Noise We must also incorporate the effect of cyclostationary noise. As explained in Section 8.7.3, such noise can be viewed as stationary noise, $n(t)$, multiplied by a periodic envelope, $e(t)$. Equation (8.138) can thus be written as

$$y(t) = \int_{-\infty}^{+\infty} n(\tau)e(\tau)h(t, \tau)d\tau, \tag{8.148}$$

implying that $e(t)h(t, \tau)$ can be viewed as an "effective" impulse response [6]. In other words, the effect of $n(t)$ on phase noise ultimately depends on the *product* of the cyclostationary noise envelope and $h(t, \tau)$.

This approach to phase noise analysis generally requires that both the noise envelope and the impulse response be determined from multiple simulations for each device. Design optimization may therefore prove a lengthy task.

Each of the two analysis approaches described thus far imparts its own insights and finds its own utility in circuit design. However, there are other phase noise mechanisms that can be better understood by other analysis techniques. The next section is an example.

8.7.5 Noise of Bias Current Source

Oscillators typically employ a bias current source so as to minimize sensitivity to the supply voltage and noise therein. We wish to study the phase noise contributed by this current source. Figure 8.79 summarizes the tail-related noise mechanisms studied here.

Consider the topology shown in Fig. 8.80(a), where I_n models the noise of I_{SS}, including flicker noise near zero frequency, thermal noise around the oscillation frequency, ω_0,

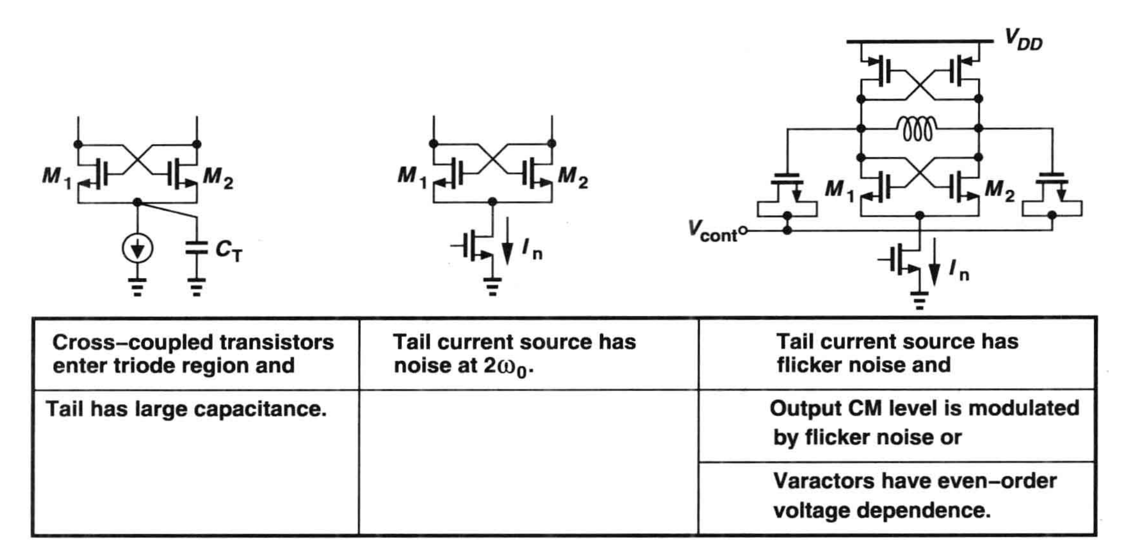


Figure 8.79 *Tail noise mechanisms in cross-coupled oscillator.*

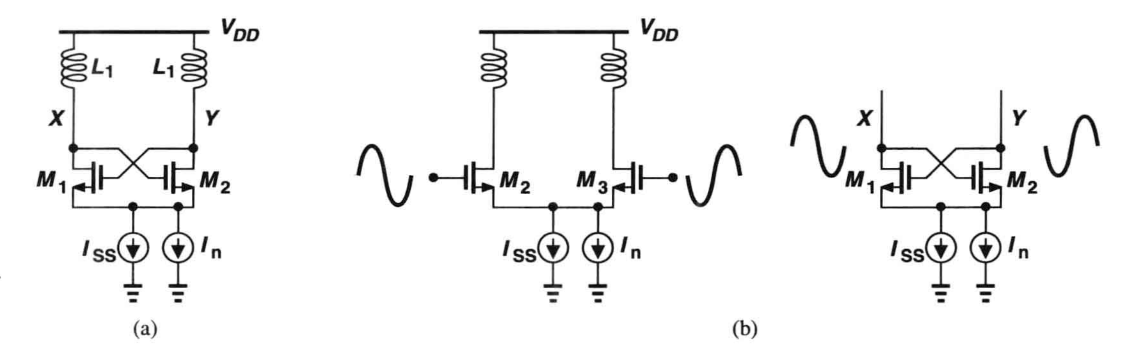


Figure 8.80 *(a) Oscillator with noisy tail current source, (b) circuit viewed as a mixer.*

thermal noise around $2\omega_0$, etc. We also recognize that M_1 and M_2 are periodically turned on and off, thus steering (commutating) $I_{SS} + I_n$ and hence operating as a *mixer*. In other words, the two circuits shown in Fig. 8.80(b) are similar, and the differential current injected by M_1 and M_2 into the tanks can be viewed as the product of $I_{SS} + I_n$ and a square wave toggling between -1 and $+1$ (for large swings).

We now examine the effect of different noise frequencies upon the performance of the oscillator in Fig. 8.80(a). The flicker noise in I_n slowly varies the bias current and hence the output voltage swing ($4I_{SS}R_p/\pi$), introducing amplitude modulation. We therefore postulate that the flicker noise produces negligible phase noise. As explained later, this is not true in the presence of voltage-dependent capacitances at the output nodes (e.g., varactors), but we neglect the effect of flicker noise for now.

How about the noise around ω_0? This noise component is mixed with the harmonics of the square wave, ω_0, $3\omega_0$, $5\omega_0$, ..., landing at 0, $2\omega_0$, $4\omega_0$, ... Thus, this component is negligible.

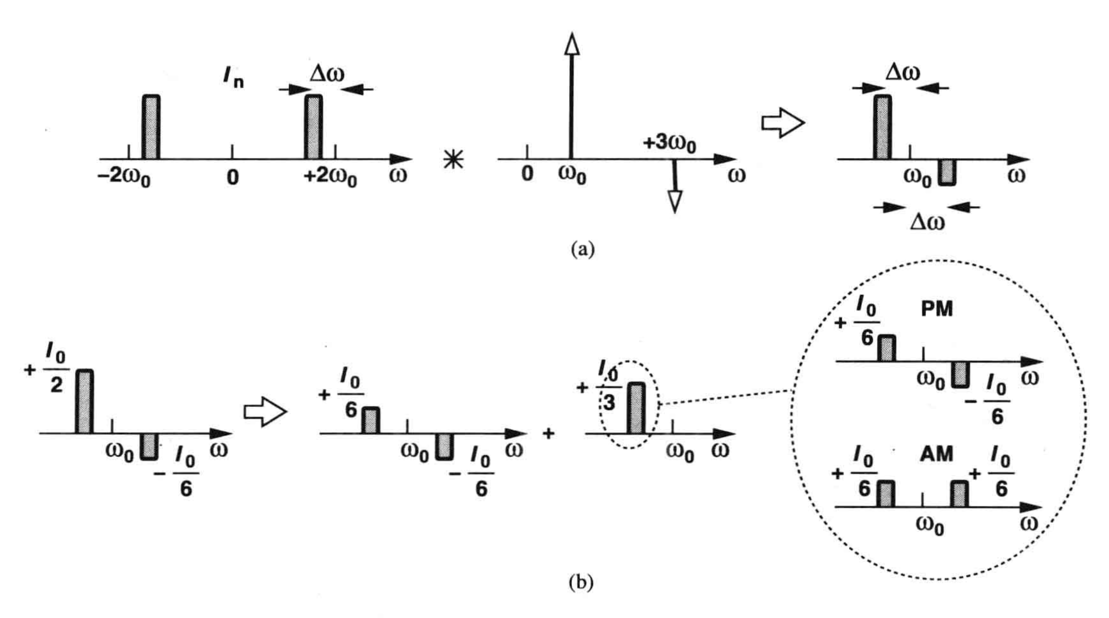


Figure 8.81 *(a) Translation of tail noise around* $2\omega_0$ *to sidebands around* ω_0*, (b) separation of PM and AM components.*

The noise around $2\omega_0$, on the other hand, markedly impacts the performance [12, 7]. As illustrated in Fig. 8.81(a), a noise component slightly below $2\omega_0$ is mixed with the first and third harmonics of the square wave, thereby falling at slightly below and above ω_0 but with *different* amplitudes and polarities. To determine whether these components produce AM or FM, we express the oscillator output as $\cos\omega_0 t$ and its third harmonic as $(-1/3)\cos(3\omega_0 t)$. For a tail current noise component, $I_0\cos(2\omega_0 - \Delta\omega)t$, the differential output current of M_1 and M_2 emerges as

$$I_{out} \propto \cos\omega_0 t I_0 \cos(2\omega_0 - \Delta\omega)t + \frac{-1}{3}\cos(3\omega_0 t) I_0 \cos(2\omega_0 - \Delta\omega)t \tag{8.149}$$

$$\propto \frac{I_0}{2}\cos(\omega_0 - \Delta\omega)t - \frac{I_0}{6}\cos(\omega_0 + \Delta\omega)t + \dots. \tag{8.150}$$

As explained in Chapter 3, two equal cosine sidebands having opposite signs surrounding a cosine carrier represent FM. In the above equation, however, the two sidebands have unequal magnitudes, creating some AM as well. Writing $(I_0/2)\cos(\omega_0 - \Delta\omega)t = (I_0/6)\cos(\omega_0 - \Delta\omega)t + (I_0/3)\cos(\omega_0 - \Delta\omega)t$ and extracting from the second term its PM components [Fig. 8.81(b)] as $(I_0/6)\cos(\omega_0 - \Delta\omega)t - (I_0/6)\cos(\omega_0 + \Delta\omega)t$, we obtain the overall PM sidebands:

$$I_{out} \propto \frac{I_0}{3}\cos(\omega_0 + \Delta\omega)t - \frac{I_0}{3}\cos(\omega_0 - \Delta\omega)t + \dots. \tag{8.151}$$

The proportionality factor is related to the conversion gain of the mixing action, as illustrated by the following example.

Example 8.37

For a tail noise of $I_n = I_0 \cos(2\omega_0 + \Delta\omega)t$ in Fig. 8.80(a), determine the magnitude of the FM sidebands in the differential output current.

Solution:

In the frequency domain, each impulse near $2\omega_0$ has a magnitude of $I_0/2$. Upon mixing with the first harmonic of the square wave, the impulse at $2\omega_0 + \Delta\omega$ appears at $\omega_0 + \Delta\omega$ with a height of $(2/\pi)(I_0/2)$. Similarly, mixing with the third harmonic yields an impulse at $\omega_0 - \Delta\omega$ with a height of $-(1/3)(2/\pi)(I_0/2)$. Separating the AM component as explained above, we have

$$I_{out} = \frac{1}{3}\frac{4}{\pi} I_0 \cos(\omega_0 + \Delta\omega)t - \frac{1}{3}\frac{4}{\pi} I_0 \cos(\omega_0 - \Delta\omega)t + \ldots . \tag{8.152}$$

To obtain the phase noise in the output *voltage*, (1) the current sidebands computed in the above example must be multiplied by the impedance of the tank at a frequency offset of $\pm\Delta\omega$, and (2) the result must be normalized to the oscillation amplitude. Note that the current components see the *lossless* impedance of the tank once they are injected into the output nodes because the average negative conductance presented by M_1 and M_2 cancels the loss. This impedance is given by $-j/(2C_1\Delta\omega)$ if $\Delta\omega \ll \omega_0$. Thus, the relative phase noise can be expressed as

$$S(\Delta\omega) = \frac{\dfrac{16\overline{I_n^2}}{9\pi^2}\left(\dfrac{1}{2C\Delta\omega}\right)^2}{\dfrac{4}{\pi^2} I_{SS}^2 R_p^2} = \frac{4\overline{I_n^2}}{9I_{SS}^2}\left(\frac{\omega_0}{2Q\Delta\omega}\right)^2. \tag{8.153}$$

The thermal noise near higher even harmonics of ω_0 plays a similar role, producing FM sidebands around ω_0. It can be shown that the summation of all of the sideband powers results in the following phase noise expression due to the tail current source [8, 10]:

$$S(\Delta\omega) = \frac{\pi^2\overline{I_n^2}}{16I_{SS}^2}\left(\frac{\omega_0}{2Q\Delta\omega}\right)^2. \tag{8.154}$$

Let us now consider the noise of the top current source in Fig. 8.30(a). We wish to formulate the frequency modulation resulting from this noise. Suppose I_{DD} contains a noise current $i_n(t)$. As calculated in Example 8.16, $i_n(t)$ produces a common-mode voltage change of

$$\Delta V = \frac{1}{g_m}\frac{i_n(t)}{2}. \tag{8.155}$$

This change is indistinguishable from an equal but opposite change in the control voltage, V_{cont}. As explained in Section 8.10, the output waveform can be expressed as

$$V_{out}(t) = V_0 \cos\left[\omega_0 t + \int K_{VCO}\frac{i_n(t)}{2g_m}dt\right], \tag{8.156}$$

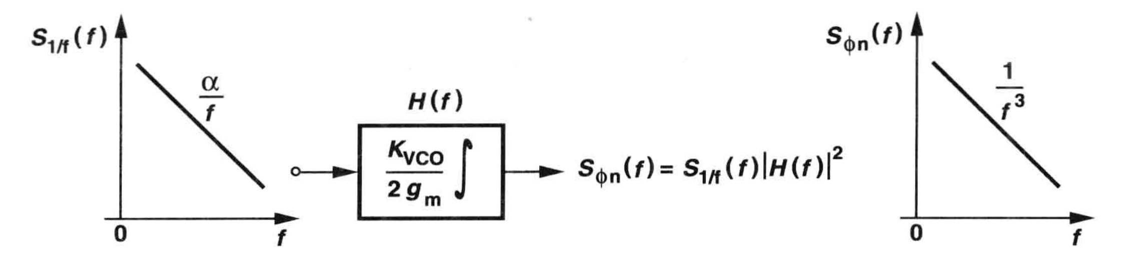


Figure 8.82 *Equivalent transfer function for conversion of top bias current to phase noise.*

where the second term in the square brackets is the resulting phase noise, $\phi_n(t)$. If $\phi_n(t) \ll 1$ rad, then

$$V_{out}(t) \approx V_0 \cos\omega_0 t - V_0 \frac{K_{VCO}}{2g_m}\left[\int i_n(t)dt\right] \sin\omega_0 t. \tag{8.157}$$

We recognize that *low-frequency* components in $i_n(t)$ are upconverted to the vicinity of ω_0. In particular, as shown in Fig. 8.82, $1/f$ noise components in $i_n(t)$ experience a transfer function of $[K_{VCO}/(2g_m)](1/s)$, producing

$$S_{\phi n}(f) = \left(\frac{K_{VCO}}{2g_m}\frac{1}{2\pi f}\right)^2 \frac{\alpha}{f}. \tag{8.158}$$

AM/PM Conversion Let us summarize our findings thus far. In the tail-biased oscillator (and in the top-biased oscillator), the noise near zero frequency introduces amplitude modulation, whereas that near even harmonics of ω_0 leads to phase noise. [In the top-biased oscillator, low-frequency noise in the current source also modulates the output CM level, producing phase noise (Example 8.16).]

The amplitude modulation resulting from the bias current noise does translate to phase noise in the presence of nonlinear capacitances in the tanks [13, 14]. To understand this point, we return to our AM/PM modulation study in Chapter 2 and make the following observations. Since the varactor capacitance varies periodically with time, it can be expressed as a Fourier series:

$$C_{var} = C_{avg} + \sum_{n=1}^{\infty} a_n \cos n\omega_0 t + \sum_{n=1}^{\infty} b_n \sin n\omega_0 t, \tag{8.159}$$

where C_{avg} denotes the "dc" value. If noise in the circuit modulates C_{avg}, then the oscillation frequency and phase are also modulated. We must therefore determine under what conditions the tail noise, i.e., the output AM noise, modulates C_{avg}.

Consider the tank shown in Fig. 8.83(a), and first assume that the voltage dependence of C_1 is odd-symmetric around the vertial axis, e.g., $C_1 = C_0(1 + \alpha V)$. In this case, C_{avg} is independent of the signal amplitude because the capacitance spends equal amounts of time above and below C_0 [Fig. 8.83(b)]. The average tank resonance frequency is thus constant and no phase modulation occurs.

The above results change if C_1 exhibits *even-order* voltage dependence, e.g., $C_1 = C_0(1 + \alpha_1 V + \alpha_2 V^2)$. Now, the capacitance changes more sharply for negative or positive voltages, yielding an average that depends on the current amplitude [Fig. 8.83(c)].

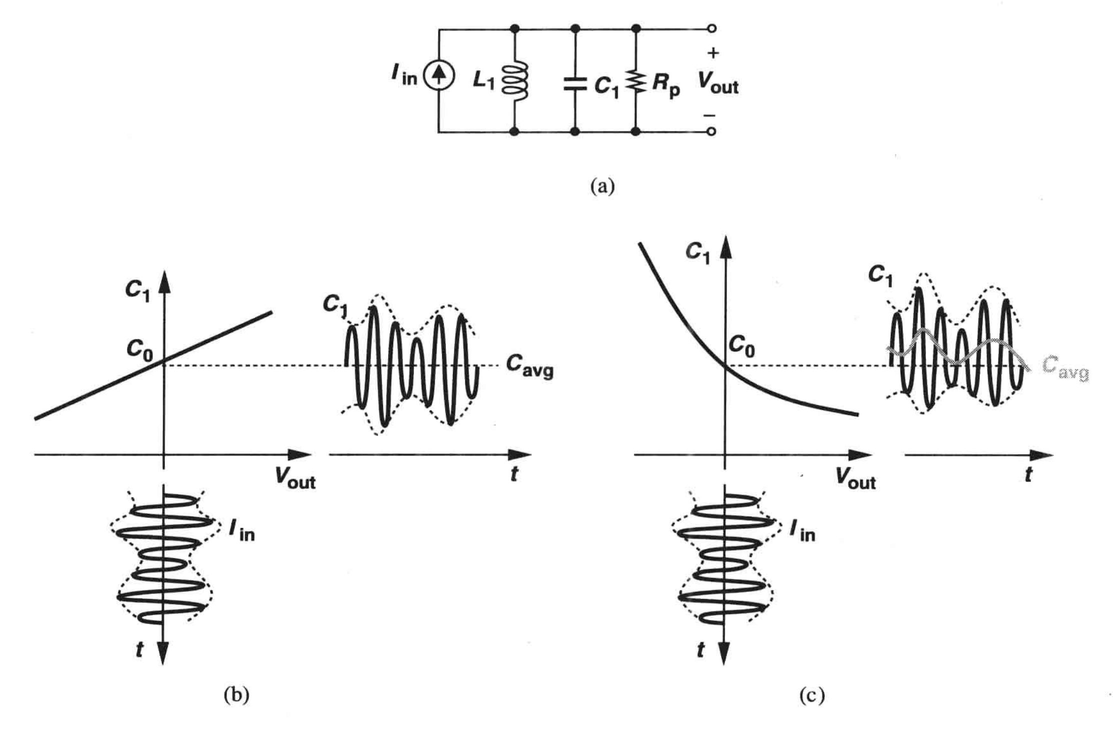


Figure 8.83 *(a) Tank driven by an RF current course, (b) effect of AM on average capacitance for a C/V characteristic that is symmetric around vertical axis, (c) effect of AM on average capacitance for a C/V characteristic that is asymmetric around vertical axis.*

We therefore observe that, in an oscillator employing such a tank, slow modulation of the amplitude varies the average tank resonance frequency and hence the frequency of oscillation. The phase noise resulting from the low-frequency bias current noise is computed in [13]. We study a simplified case in Problem 8.13.

It is important to note that the tail current in Fig. 8.35 introduces phase noise via three distinct mechanisms: (1) its flicker noise modulates the output CM level and hence the varactors; (2) its flicker noise produces AM at the output and hence phase noise through AM/PM conversion; (3) its thermal noise at $2\omega_0$ gives rise to phase noise.

8.7.6 Figures of Merit of VCOs

Our studies in this chapter point to direct trade-offs among the phase noise, power dissipation, and tuning range of VCOs. For example, as explained in Chapter 7, varactors themselves suffer from a trade-off between their capacitance range and their Q. We also recall from Leeson's equation, Eq. (8.107), that phase noise rises with the oscillation frequency if the Q does not increase proportionally.

A figure of merit (FOM) that encapsulates some of these trade-offs is defined as

$$\text{FOM}_1 = \frac{(\text{Oscillation Frequency})^2}{\text{Power Dissipation} \times \text{Phase Noise} \times (\text{Offset Frequency})^2}, \tag{8.160}$$

where the phase noise is multiplied by the square of the offset frequency at which it is measured so as to perform normalization. Attention must be paid to the unit of phase noise (noise power normalized to carrier power) in this expression. Note that the product of power dissipation and phase noise has a unit of W/Hz. Another FOM that additionally represents the trade-offs with the tuning range is

$$\mathrm{FOM}_2 = \frac{(\text{Oscillation Frequency})^2}{\text{Power Dissipation} \times \text{Phase Noise} \times (\text{Offset Frequency})^2} \times \left(\frac{\text{Tuning Range}}{\text{Oscillation Frequency}}\right)^2. \tag{8.161}$$

State-of-the-art CMOS VCOs in the range of several gigahertz achieve an FOM_2 around 190 dB. In general, the phase noise in the above expressions refers to the worst-case value, typically at the the highest oscillation frequency. Also, note that these FOMs do not account for the load driven by the VCO.

8.8 DESIGN PROCEDURE

Following our study of tuning techniques and phase noise issues, we now describe a procedure for the design of LC VCOs. We focus on the topology shown in Fig. 8.25(a) and assume the following parameters are given: the center frequency, ω_0, the output voltage swing, the power dissipation, and the load capacitance, C_L. Even though some of these parameters may not be known at the outset, it is helpful to select some reasonable values and iterate if necessary. Of course, the output swing must be chosen so as not to stress the transistors.

The procedure consists of six steps:

1. Based on the power budget and hence the maximum allowable I_{SS}, select the tank parallel resistance, R_p, so as to obtain the required voltage swing, $(4/\pi)I_{SS}R_p$.
2. Select the *smallest* inductor value that yields a parallel resistance of R_p at ω_0, i.e., find the inductor with the maximum $Q = R_p/(L\omega_0)$. This, of course, relies on detailed modeling and characterization of inductors in the technology at hand (Chapter 7). Denote the capacitance contributed by the inductors to each node by C_p.
3. Determine the dimensions of M_1 and M_2 such that they experience nearly complete switching with the given voltage swings. To minimize their capacitance contributions, choose minimum channel length for the transistors.
4. Noting that the transistor, inductor, and load capacitances amount to a total of $C_{GS} + 4C_{GD} + C_{DB} + C_p + C_L$ at each output node, calculate the *maximum* varactor capacitance, $C_{var,max}$, that can be added to reach the lower end of the tuning range, ω_{min}; e.g., $\omega_{min} = 0.9\omega_0$:

$$\frac{1}{\sqrt{L_0(C_{GS} + 4C_{GD} + C_{DB} + C_p + C_L + C_{var,max})}} \approx 0.9\omega_0. \tag{8.162}$$

5. Using proper varactor models, determine the minimum capacitance of such a varactor, $C_{var,min}$, and compute the upper end of the tuning range,

$$\omega_{max} = \frac{1}{\sqrt{L_0(C_{GS} + 4C_{GD} + C_{DB} + C_p + C_L + C_{var,min})}}. \tag{8.163}$$

6. If ω_{max} is quite higher than necessary, increase $C_{var,max}$ in Eq. (8.162) so as to center the tuning range around ω_0.

The above procedure yields a design that achieves the *maximum* tuning range subject to known values of ω_0, the output swing, the power dissipation, and C_L. If the varactor Q is high enough, such a design also has the highest tank Q. At this point, we calculate or simulate the phase noise for different frequencies across the tuning range. If the phase noise rises significantly at ω_{min} or ω_{max}, then the tuning range must be reduced, and if it is still excessively high, the design procedure must be repeated with a higher power budget. In applications requiring a low phase noise, a multitude of different VCO topologies must be designed and simulated so as to obtain a solution with an acceptable performance.

We carry out the detailed design of a 12-GHz VCO for 11a/g applications in Chapter 13.

Example 8.38

If the power budget allocated to the VCO topology of Fig. 8.25(a) is doubled, by what factor is the phase noise reduced?

Solution:

Doubling the power budget can be viewed as (a) placing two identical oscillators in parallel [Fig. 8.84(a)] or (b) scaling all of the components in an oscillator by a factor of 2 [Fig. 8.84(b)]. In this scenario, the output voltage swing and the tuning range remain unchanged (why?), but the phase noise power falls by a factor of two (3 dB). This is because in Eq. (8.124), R_p is doubled and I_{SS}^2 is quadrupled.

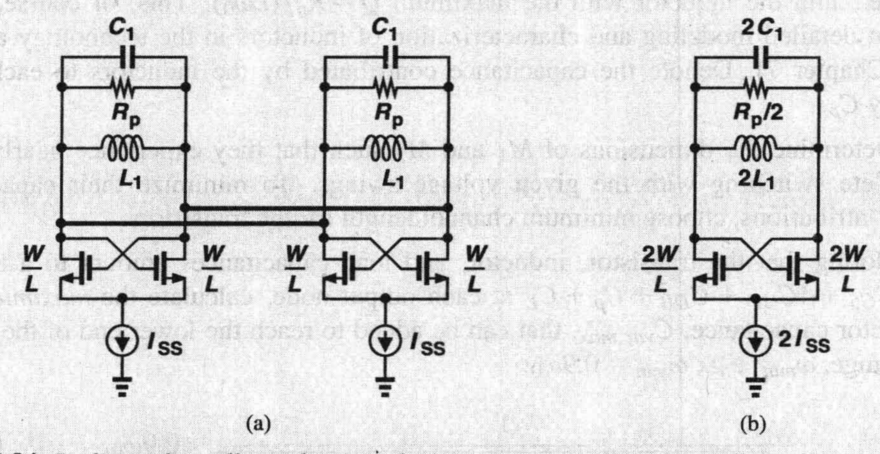


Figure 8.84 *Scaling of oscillator by (a) placing two instances in parallel, (b) scaling each component.*

8.8.1 Low-Noise VCOs

A great deal of effort has been expended on relaxing the trade-offs among phase noise, power dissipation, and tuning range of VCOs. In this section, we study several examples.

In order to reduce the phase noise due to flicker noise, the generic cross-coupled oscillator can incorporate PMOS transistors rather than NMOS devices. Figure 8.85 depicts the PMOS counterparts of the circuits shown in Figs. 8.25(a) and 8.30(a). Since PMOS devices exhibit substantially less flicker noise, the close-in phase noise of these oscillators is typically 5 to 10 dB lower. The principal drawback of these topologies is their limited speed, an issue that arises only as frequencies exceeding tens of gigahertz are sought.

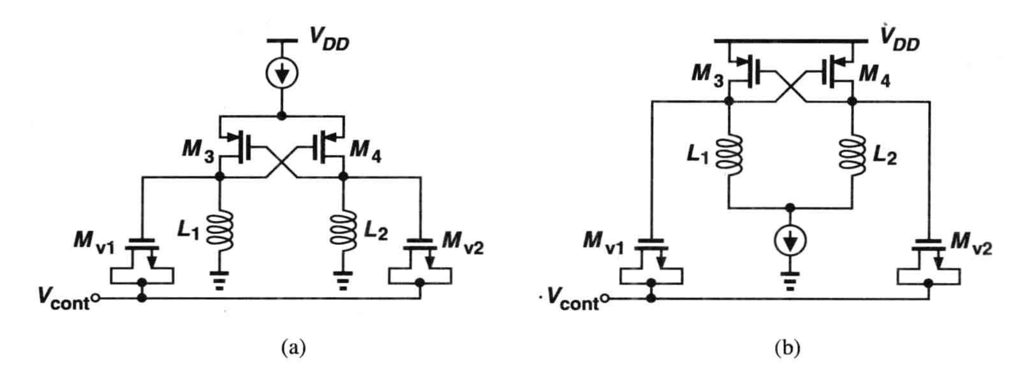


Figure 8.85 *(a) Tail-biased and (b) bottom-biased PMOS oscillators.*

As explained in Section 8.7.5, the noise current at $2\omega_0$ in the tail current source translates to phase noise around ω_0. This and higher noise harmonics can be removed by a capacitor as shown in Fig. 8.86(a). However, if M_1 and M_2 enter the deep triode region during oscillation, then two effects raise the phase noise: (1) the on-resistance of each transistor now degrades the Q of the tank [Fig. 8.69(a)] [16], and (2) the impulse response (ISF) from the noise of each transistor to the output phase becomes substantially larger [17]. This

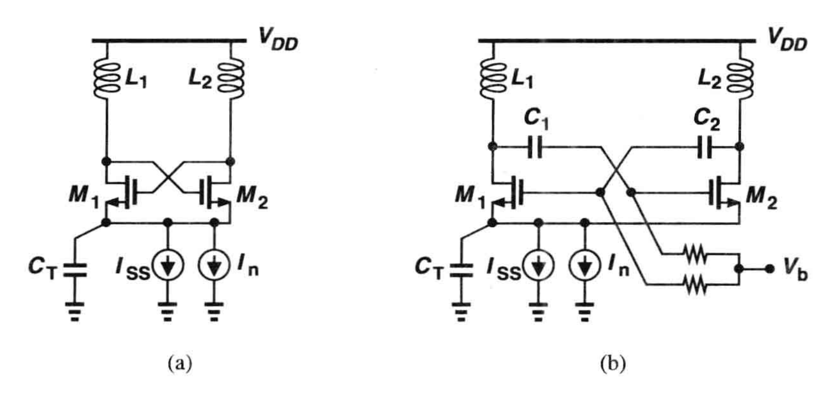


Figure 8.86 *(a) Use of capacitor to shunt tail noise, (b) use of ac coupling to avoid operation in deep triode region.*

issue can be resolved by means of two techniques. If operation in the triode region must be avoided but large output swings are desired, capacitive coupling can be inserted in the loop [Fig. 8.86(b)]. Here, V_b is chosen so that the peak voltage at the gate of each transistor does not exceed its minimum drain voltage by more than one threshold. A Class-C oscillator similar to this topology is described in [17].

Example 8.39

If C_1 and C_2 along with transistor capacitances attenuate the swing by a factor of 2, determine the requisite value of V_b so that the transistors are in saturation.

Solution:

Illustrated in Fig. 8.87 are the gate and drain waveforms. For the transistor to remain in saturation,

$$\frac{V_p}{2} + V_b - (V_{DD} - V_p) \leq V_{TH} \tag{8.164}$$

and hence

$$V_b \leq V_{DD} - \frac{3V_p}{2} + V_{TH}. \tag{8.165}$$

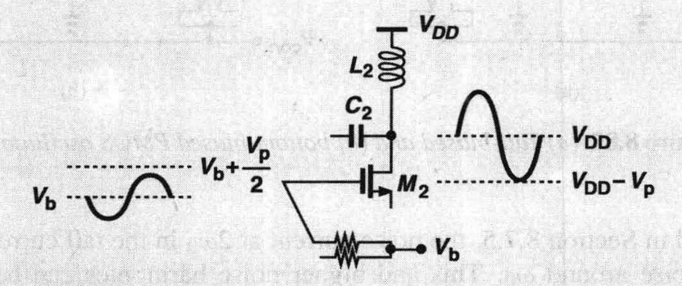


Figure 8.87 *Gate and drain swings with capacitive coupling.*

The bias voltage, V_b, in Fig. 8.86(b) must remain high enough to provide sufficient V_{GS} for the cross-coupled transistors and headroom for the tail current source. Consequently, the output swings may still be severely limited. In Problem 8.14, we show that the peak swing does not exceed approximately $V_{DD} - 2(V_{GS} - V_{TH})$. This is obtained only if the capacitive attenuation is so large as to yield a negligible gate swing, requiring a high g_m.

The second approach is to allow M_1 and M_2 in Fig. 8.86(a) to enter the triode region but remove the effect of the tail capacitance at $2\omega_0$. Illustrated in Fig. 8.88 [16], the idea is to insert inductor L_T in series with the tail node and choose its value such that it resonates with the parasitic capacitance, C_B, at $2\omega_0$. The advantage of this topology over that in Fig. 8.86(b) is that it affords larger swings. The disadvantage is that it employs an additional inductor and requires tail tuning for broadband operation.

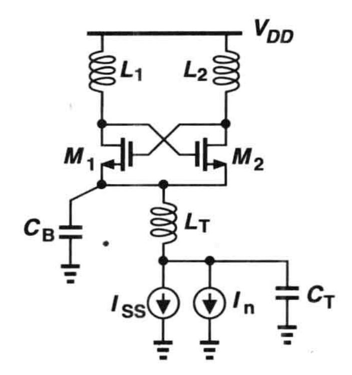


Figure 8.88 *Use of tail resonance to avoid tank Q degradation in deep triode region.*

Example 8.40

Study the behavior of the circuit shown in Fig. 8.88 if the supply voltage contains high-frequency noise.

Solution:

Capacitor C_B degrades the high-frequency common-mode rejection of the circuit. This issue can be partially resolved by tying C_B to V_{DD} (Fig. 8.89). Now, C_B bootstraps node P to V_{DD} at high frequencies. Of course, this is not possible if C_B arises from only the parasitics at the tail node.

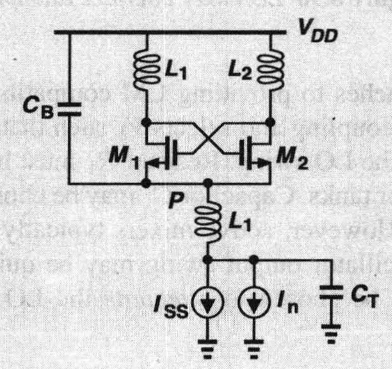


Figure 8.89 *Connection of capacitor to V_{DD} to improve supply rejection.*

8.9 LO INTERFACE

Each oscillator in an RF system typically drives a mixer and a frequency divider, experiencing their input capacitances. Moreover, the LO output common-mode level must be compatible with the input CM level of these circuits. We study this compatibility issue here.

The output CM level of the oscillators studied in this chapter is around V_{DD}, $V_{DD}/2$, or zero [for the PMOS implementation of Fig. 8.85(a)]. On the other hand, the required

input CM level of the mixers studied in Chapter 6 is somewhat higher than $V_{DD}/2$ for active NMOS topologies or around V_{DD} (zero) for passive NMOS (PMOS) realizations. Figure 8.90 illustrates some of the LO/mixer combinations, suggesting that dc coupling is possible in only some cases.

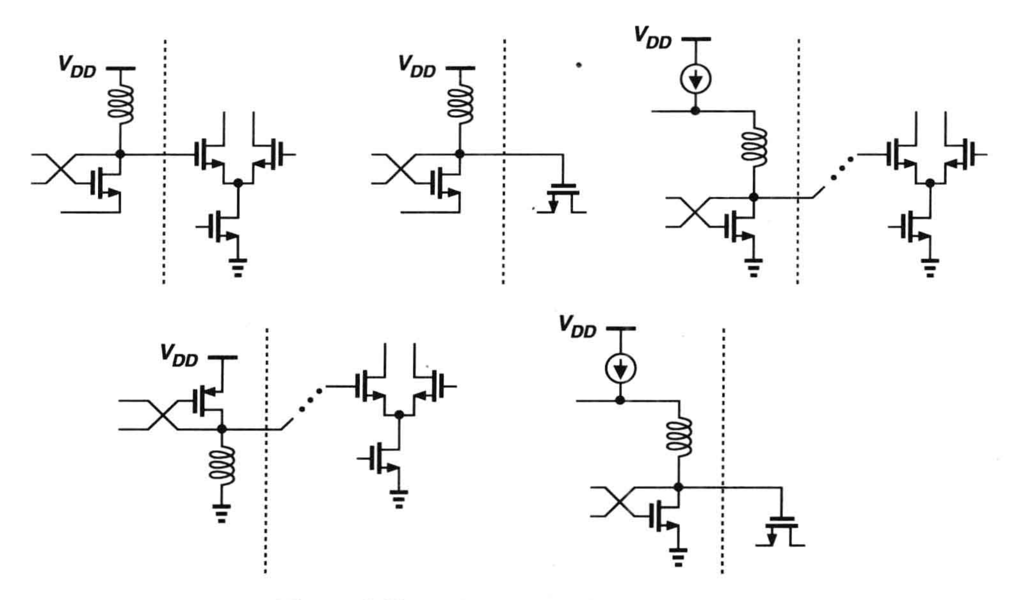


Figure 8.90 *LO/mixer interface examples.*

We consider two approaches to providing CM compatibility. Shown in Fig. 8.91(a), the first employs capacitive coupling and selects V_b such that M_1 and M_2 do not enter the triode region at the peak of the LO swing. Resistor R_1 must be large enough to negligibly degrade the Q of the oscillator tanks. Capacitor C_1 may be chosen 5 to 10 times C_{in} to avoid significant LO attenuation. However, active mixers typically operate with only moderate LO swings, whereas the oscillator output swing may be quite larger so as to reduce its phase noise. Thus, C_1 may be chosen to *attenuate* the LO amplitude. For example, if

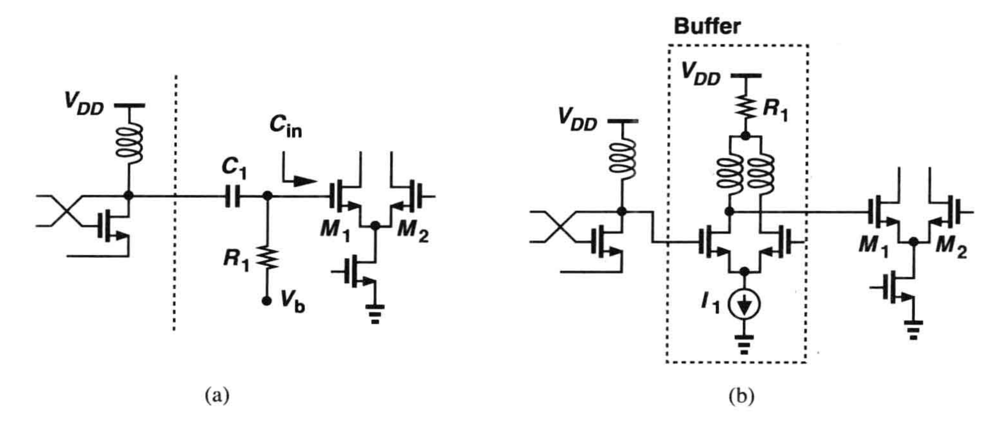


Figure 8.91 *Use of (a) capacitive coupling, (b) a buffer between LO and mixer.*

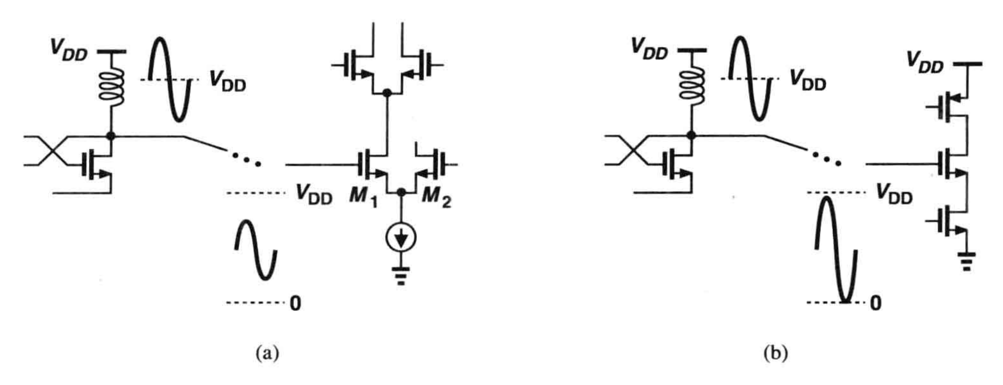


Figure 8.92 *LO/divider interface for (a) current-steering and (b) rail-to-rail operation.*

$C_1 = C_{in}$, an attenuation factor of 2 results *and* the capacitance presented to the LO falls to $C_1C_{in}/(C_1 + C_{in}) = C_{in}/2$, a useful attribute afforded by active mixers.

The second approach to CM compatibility interposes a buffer between the LO and the mixer. In fact, if the mixer input capacitance excessively loads the LO, or if long, lossy interconnects appear in the layout between the LO and the mixer, then such a buffer proves indispensable. Depicted in Fig. 8.91(b) is an example, where an inductively-loaded differential pair serves as a buffer, providing an output CM level equal to $V_{DD} - I_1R_1$. This value can be chosen to suit the LO port of the mixer. The drawback of this approach stems from the use of additional inductors and the resulting routing complexity.

Similar considerations apply to the interface between an oscillator and a frequency divider. For example, in Fig. 8.92(a), the divider input CM level must be well below V_{DD} to ensure the current-steering transistors M_1 and M_2 do not enter the deep triode region (Chapter 10). As another example, some dividers require a rail-to-rail input [Fig. 8.92(b)], and possibly capacitive coupling.

8.10 MATHEMATICAL MODEL OF VCOs

Our definition of voltage-controlled oscillators in Section 8.5 relates the output frequency to the control voltage by a linear, static equation, $\omega_{out} = \omega_0 + K_{VCO}V_{cont}$. But, how do we express the output in the time domain? To this end, we must reexamine our understanding of frequency and phase.

Example 8.41

Plot the waveforms for $V_1(t) = V_0 \sin \omega_1 t$ and $V_2(t) = V_0 \sin(at^2)$.

Solution:

To plot these waveforms carefully, we must determine the time instants at which the argument of the sine reaches integer multiples of π. For $V_1(t)$, the argument, $\omega_1 t$, rises linearly

(Continues)

Example 8.41 (*Continued*)

with time, crossing $k\pi$ at $t = \pi k/\omega_1$ [Fig. 8.93(a)]. For $V_2(t)$, on the other hand, the argument rises increasingly faster with time, crossing $k\pi$ more frequently. Thus, $V_2(t)$ appears as shown in Fig. 8.93(b).

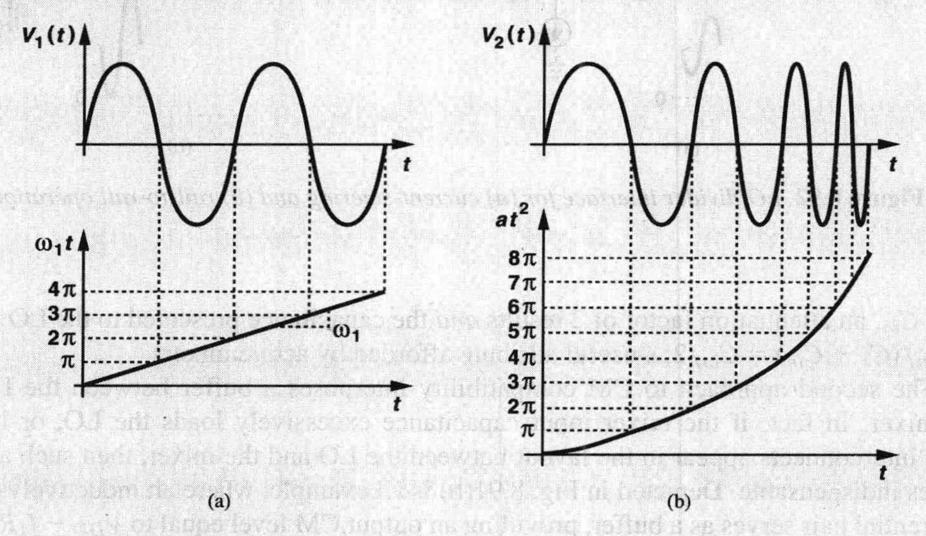


Figure 8.93 *(a) Linear and (b) quadratic growth of phase with time.*

Example 8.42

Since a sinusoid of constant frequency ω_1 can be expressed as $V_0 \cos \omega_1 t$, a student surmises that the output waveform of a VCO can be written as

$$V_{out}(t) = V_0 \cos \omega_{out} t \tag{8.166}$$

$$= V_0 \cos(\omega_0 + K_{VCO} V_{cont})t. \tag{8.167}$$

Explain why this is incorrect.

Solution:

As an example, suppose $V_{cont} = V_m \sin \omega_m t$, i.e., the frequency of the oscillator is modulated periodically. Intuitively, we expect the output waveform shown in Fig. 8.94(a), where the frequency periodically swings between $\omega_0 + K_{VCO}V_m$ and $\omega_0 - K_{VCO}V_m$, i.e., has a "peak deviation" of $\pm K_{VCO}V_m$. However, the student's expression yields

$$V_{out}(t) = V_0 \cos[\omega_0 t + K_{VCO} V_m (\sin \omega_m t)t]. \tag{8.168}$$

Example 8.42 *(Continued)*

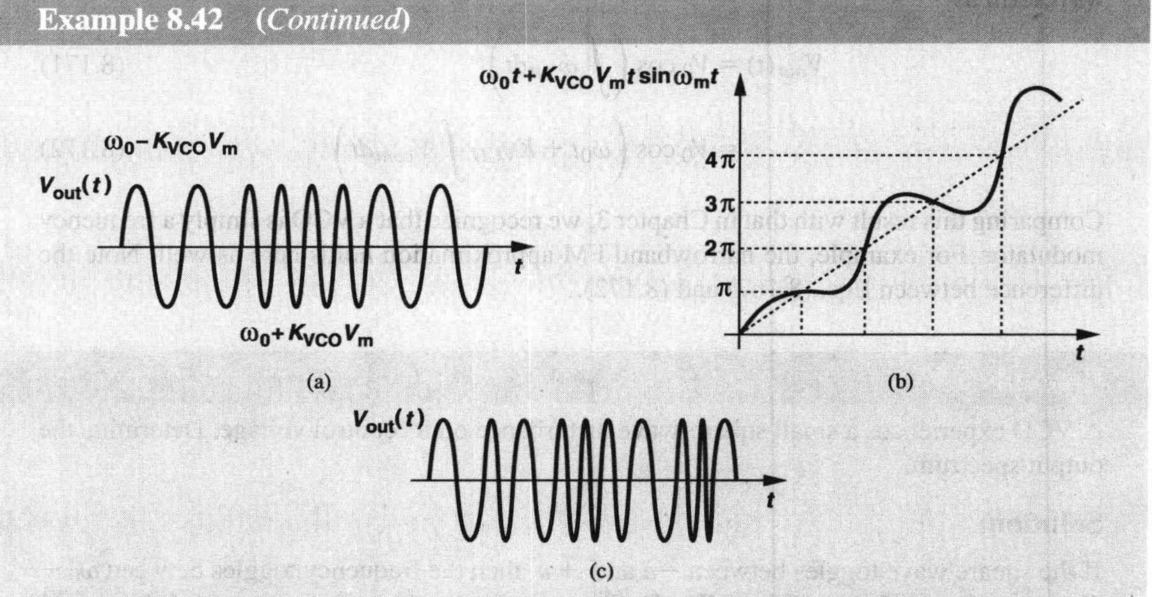


Figure 8.94 *(a) Frequency-modulated sinusoid, (b) incorrect definition of phase, (c) corresponding output waveform.*

Let us plot this waveform, noting from Example 8.41 that we must identify the time instants at which the argument crosses integer multiples of π. Recognizing that the second term of the argument displays a growing amplitude, we plot the overall argument as depicted in Fig. 8.94(b) and draw horizontal lines corresponding to $k\pi$. The intersection of each horizontal line with the phase plot signifies the zero crossings of $V_{out}(t)$. Thus, $V_{out}(t)$ appears as shown in Fig. 8.94(c). The key point here is that the VCO frequency is not modulated periodically.

Let us now consider an unmodulated sinusoid, $V_1(t) = V_0 \sin \omega_1 t$. Called the "total phase," the argument of the sine, $\omega_1 t$, varies linearly with time in this case, exhibiting a slope of ω_1 [Fig. 8.93(a)]. We say the phase "accumulates" at a rate of ω_1. In other words, if ω_1 is increased to ω_2, then the phase accumulates faster, crossing multiples of π at a higher rate. It is therefore plausible to define the instantaneous frequency as the time derivative of the phase:

$$\omega = \frac{d\phi}{dt}. \tag{8.169}$$

Conversely,

$$\phi = \int \omega dt + \phi_0. \tag{8.170}$$

The initial phase, ϕ_0, is usually unimportant and assumed zero hereafter. Since a VCO exhibits an output frequency given by $\omega_0 + K_{VCO} V_{cont}$, we can express its output

waveform as

$$V_{out}(t) = V_0 \cos\left(\int \omega_{out} dt\right) \tag{8.171}$$

$$= V_0 \cos\left(\omega_0 t + K_{VCO} \int V_{cont} dt\right). \tag{8.172}$$

Comparing this result with that in Chapter 3, we recognize that a VCO is simply a frequency modulator. For example, the narrowband FM approximation holds here as well. Note the difference between Eqs. (8.167) and (8.172).

Example 8.43

A VCO experiences a small square-wave disturbance on its control voltage. Determine the output spectrum.

Solution:

If the square wave toggles between $-a$ and $+a$, then the frequency toggles between $\omega_0 - K_{VCO}a$ and $\omega_0 + K_{VCO}a$ [Fig. 8.95(a)]. To compute the spectrum, we expand the square wave in its Fourier series,

$$V_{cont}(t) = a\left(\frac{4}{\pi}\cos\omega_m t - \frac{1}{3}\frac{4}{\pi}\cos 3\omega_m t + \cdots\right), \tag{8.173}$$

and hence

$$V_{out}(t) = V_0 \cos\left[\omega_0 t - K_{VCO}a\left(\frac{1}{\omega_m}\frac{4}{\pi}\sin\omega_m t - \frac{1}{9\omega_m}\frac{4}{\pi}\sin 3\omega_m t + \cdots\right)\right]. \tag{8.174}$$

If $4K_{VCO}a/(\pi\omega_m) \ll 1$ rad, then the narrowband FM approximation applies:

$$V_{out}(t) \approx V_0 \cos\omega_0 t + \left[K_{VCO}a\left(\frac{1}{\omega_m}\frac{4}{\pi}\sin\omega_m t - \frac{1}{9\omega_m}\frac{4}{\pi}\sin 3\omega_m t + \cdots\right)\right] V_0 \sin\omega_0 t. \tag{8.175}$$

Depicted in Fig. 8.95(b), the spectrum consists of the carrier at ω_0 and sidebands at $\omega_0 \pm \omega_m$, $\omega_0 \pm 3\omega_m$, etc.

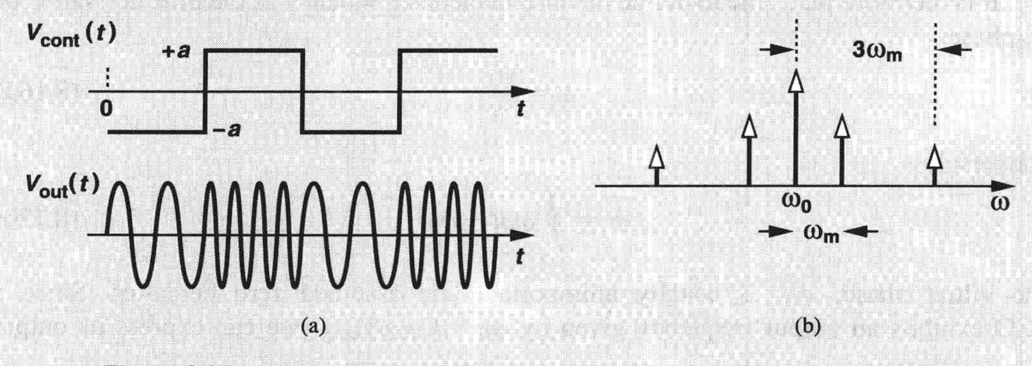


Figure 8.95 *(a) Frequency modulation by a square wave, (b) resulting spectrum.*

In the analysis of phase-locked frequency synthesizers (Chapter 10), we are concerned with only the second term in the argument of Eq. (8.172). Called the "excess phase," this term represents an *integrator* behavior for the VCO. In other words, if the quantity of interest at the output of the VCO is the excess phase, ϕ_{ex}, then

$$\phi_{ex} = K_{VCO} \int V_{cont} dt \tag{8.176}$$

and hence

$$\frac{\phi_{ex}}{V_{cont}}(s) = \frac{K_{VCO}}{s}. \tag{8.177}$$

The important observation here is that the output *frequency* of a VCO (almost) instantaneously changes in response to a change in V_{cont}, whereas the output *phase* of a VCO takes time to change and "remembers" the past.

8.11 QUADRATURE OSCILLATORS

In our study of transceiver architectures in Chapter 4, we observed the need for quadrature LO phases in downconversion and upconversion operations. We also noted that flipflop-based divide-by-two circuits generate quadrature phases, but they restrict the maximum LO frequency. In applications where dividers do not offer sufficient speed, we may employ polyphase filters or quadrature oscillators instead. In this section, we study the latter.

8.11.1 Basic Concepts

Two identical oscillators can be "coupled" such that they operate in-quadrature. We therefore begin our study with the concept of coupling (or injecting) a signal to an oscillator. Figure 8.96 depicts an example, where the input voltage is converted to current and injected into the oscillator. The differential pair is a natural means of coupling because the cross-coupled pair can also be viewed as a circuit that steers and injects current into the tanks. If the two pairs completely steer their respective tail currents, then the "coupling factor" is equal to I_1/I_{SS}.[10] This topology also exemplifies "unilateral" coupling because very little

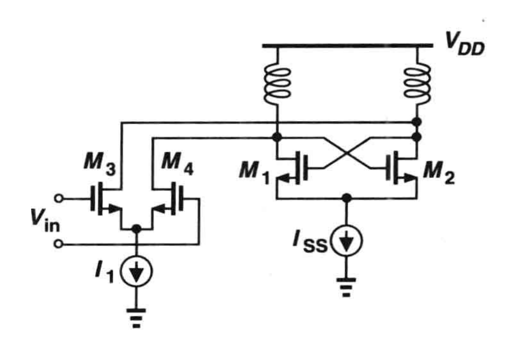


Figure 8.96 *Unilateral injection into an oscillator.*

10. In this circuit, we typically scale the transistor widths in proportion to their bias currents; thus, $g_{m3,4}/g_{m1,2} = I_1/I_{SS}$. For small-signal analysis, the coupling factor is equal to $g_{m3,4}/g_{m1,2}$.

of the oscillator signal couples back to the input. By contrast, if implemented by, say, two capacitors tied between V_{in} and the oscillator nodes, the coupling is bilateral.

Let us now consider two identical oscillators that are unilaterally coupled. Shown in Fig. 8.97 are two possibilities, with "in-phase" and "anti-phase" coupling. The coupling factors have the same sign in the former and opposite signs in the latter. We analyze these topologies using both the feedback model and the one-port model of the oscillators. Note that the tuning techniques described earlier in this chapter apply to these topologies as well.

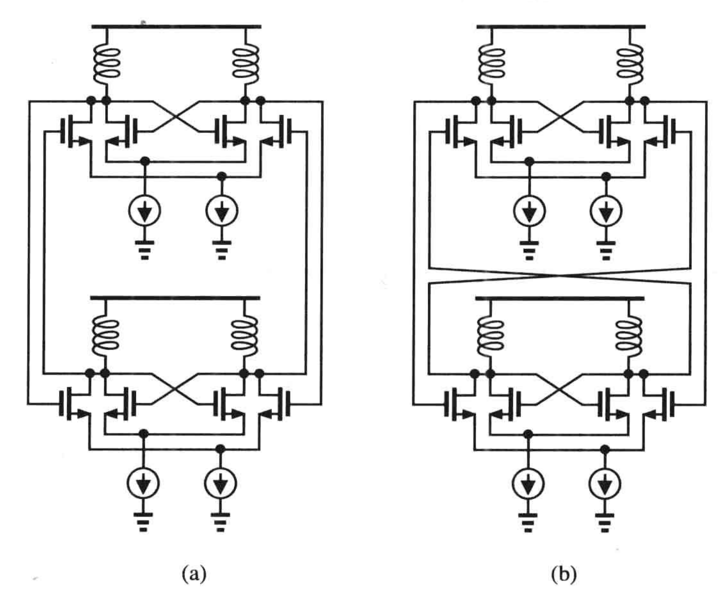


Figure 8.97 *In-phase and anti-phase coupling of two oscillators.*

Feedback Model The circuits of Fig. 8.97 can be mapped to two coupled feedback oscillators as shown in Fig. 8.98, where $|\alpha_1| = |\alpha_2|$ and the sign of $\alpha_1\alpha_2$ determines in-phase or anti-phase coupling [18]. The output of the top adder is equal to $\alpha_1 Y - X$, yielding



$$X = (\alpha_1 Y - X)H(s). \quad (8.178)$$

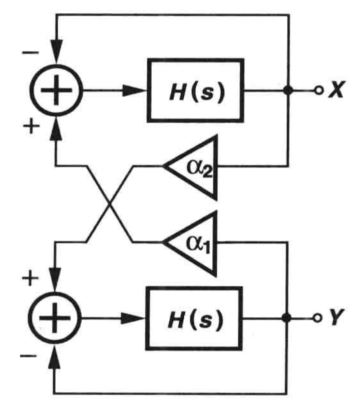


Figure 8.98 *Feedback model of quadrature oscillator.*

Similarly, the bottom oscillator produces

$$Y = (\alpha_2 X - Y)H(s). \tag{8.179}$$

Multiplying both sides of (8.178) by $\alpha_2 X$ and both sides of (8.179) by $\alpha_1 Y$ and subtracting the results, we have

$$\left(\alpha_2 X^2 - \alpha_1 Y^2\right)[1 + H(s)] = 0. \tag{8.180}$$

As explained below, $1 + H(s) \neq 0$ at the oscillation frequency, and hence

$$\alpha_2 X^2 = \alpha_1 Y^2. \tag{8.181}$$

If $\alpha_1 = \alpha_2$ (in-phase coupling), then $X = \pm Y$, i.e., the two oscillators operate with a zero or 180° phase difference.

Example 8.44

Applying Barkhausen's criteria, explain why $1 + H(s) \neq 0$ at the oscillation frequency if $\alpha_1 = \alpha_2$.

Solution:

Since each oscillator receives an additional input from the other, the oscillation startup condition must be revisited. Drawing one half of the circuit as shown in Fig. 8.99(a), we note that the input path can be merged with the feedback path as illustrated in Fig. 8.99(b). The equivalent loop transmission is therefore equal to $-(1 \pm \alpha_1)H(s)$, which, according to Barkhausen, must be equal to unity:

$$-(1 \pm \alpha_1)H(s = j\omega_0) = 1. \tag{8.182}$$

That is,

$$H(j\omega_0) = \frac{-1}{1 \pm \alpha_1}. \tag{8.183}$$

This equation serves as the startup condition.

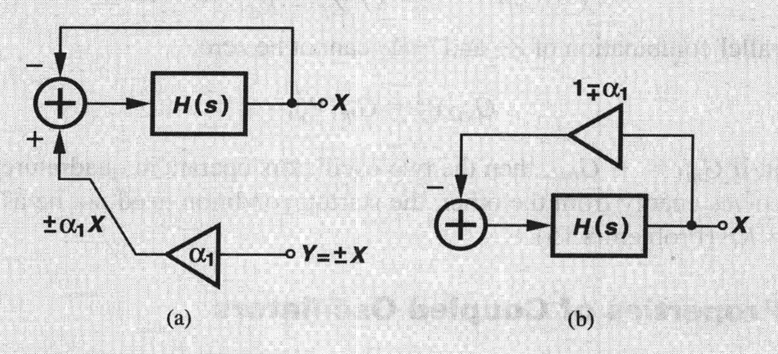


Figure 8.99 *(a) Simplified model of coupled oscillators, (b) equivalent system.*

The in-phase operation is not particularly useful as the two oscillators can be simply merged into one (as in Fig. 8.84). On the other hand, if $\alpha_1 = -\alpha_2$ (anti-phase coupling), then Eq. (8.181) yields

$$X = \pm jY; \tag{8.184}$$

i.e., the outputs bear a phase difference of $+90^\circ$ or -90°. In this case, Eq. (8.182) is revised to

$$-(1 \pm j\alpha_1)H(s = j\omega_0) = 1. \tag{8.185}$$

One-Port Model It is possible to gain additional insight through the use of the one-port model for each oscillator. A single oscillator experiencing unilateral coupling can be represented as shown in Fig. 8.100(a), where G_m denotes the transconductance of the coupling differential pair (M_3 and M_4 in Fig. 8.96), Z_T the tank impedance, and $-R_C$ the negative resistance provided by the cross-coupled pair. Two identical coupled oscillators are therefore modeled as depicted in Fig. 8.100(b), where the sign of $G_{m1}G_{m2}$ determines in-phase or anti-phase coupling.

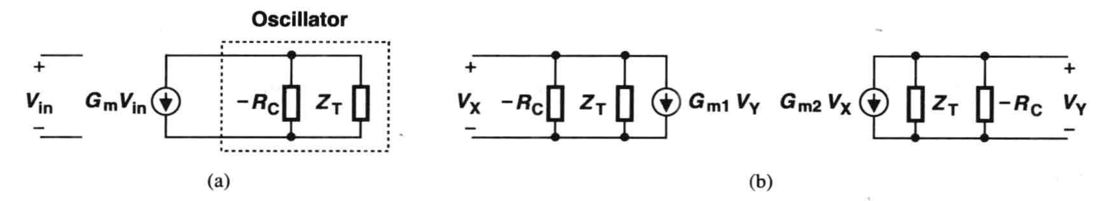


Figure 8.100 *(a) One-port oscillator model with injection, (b) model of coupled oscillators.*

The parallel combination of Z_T and $-R_C$ is given by $-Z_TR_C/(Z_T - R_C)$, yielding

$$G_{m2}V_X\frac{Z_TR_C}{Z_T - R_C} = V_Y \tag{8.186}$$

$$G_{m1}V_Y\frac{Z_TR_C}{Z_T - R_C} = V_X. \tag{8.187}$$

Multiplying both sides of (8.186) by V_X and both sides of (8.187) by V_Y and subtracting the results, we have

$$\left(G_{m2}V_X^2 - G_{m1}V_Y^2\right)\frac{Z_TR_C}{Z_T - R_C} = 0. \tag{8.188}$$

Since the parallel combination of Z_T and $-R_C$ cannot be zero,

$$G_{m2}V_X^2 = G_{m1}V_Y^2, \tag{8.189}$$

implying that, if $G_{m1} = -G_{m2}$, then the two oscillators operate in quadrature. Since each oscillator receives energy from the other, the startup condition need not be as stringent as $Z_T(s = \omega_0) = R_C$ (Problem 8.15).

8.11.2 Properties of Coupled Oscillators

Unilaterally-coupled oscillators exhibit interesting attributes. Let us consider the case of in-phase coupling as a starting point. As shown in Fig. 8.101, we construct a phasor diagram

of the circuit's voltages and currents. This is accomplished by noting that (1) V_A and V_B are 180° out of phase, and so are V_C and V_D, and (2) the drain current of each transistor is aligned with its gate voltage phasor. Thus, the total current flowing through Z_A is equal to the sum of the I_{D1} and I_{D3} phasors, altering the startup condition according to Eq. (8.182).

Can the circuit operate with I_{D3} *opposing* I_{D1}, i.e., with $V_C = -V_B$ and $V_D = -V_A$? In other words, which one of the solutions, $X = +Y$ or $X = -Y$ in Eq. (8.181), does the circuit select? From Eq. (8.182), $V_C = -V_B$ is equivalent to a lower loop gain and hence a more slowly growing amplitude. The circuit prefers to begin with the I_{D3} *enhancing* I_{D1} (Fig. 8.101), but this phase ambiguity may exist.

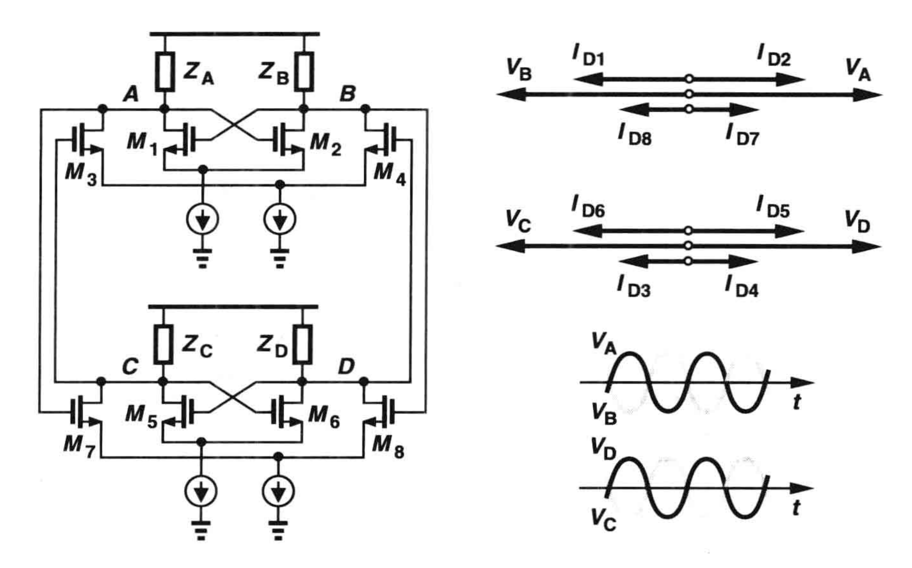


Figure 8.101 *Phasor diagrams for in-phase coupling.*

We now repeat this study for anti-phase coupling. Since the two differential oscillators operate in quadrature, the voltage and current phasors appear as in Fig. 8.102, with the drain current phasor of each transistor still aligned with its gate voltage phasor. In this case, the total current flowing through each tank consists of two *orthogonal* phasors; e.g., Z_A carries I_{D1} and I_{D3}.

How can the vector sum $I_{D1} + I_{D3}$ yield V_A? Depicted in Fig. 8.103(a), the resultant, I_{ZA}, bears an angle of θ with respect to I_{D1}. Thus, the tank must *rotate* I_{ZA} clockwise by an amount equal to θ as it converts the current to voltage. In other words, the tank impedance must provide a phase shift of θ. This is possible only if the oscillation frequency departs from the resonance frequency of the tanks. As illustrated in Fig. 8.103(b), oscillation occurs at a frequency, ω_{osc1}, such that the tank phase shift reaches θ. From Fig. 8.103(a), the tank must rotate I_{ZA} by an amount equal to

$$\angle Z_A = -\tan^{-1}\frac{I_{D3}}{I_{D1}}, \tag{8.190}$$

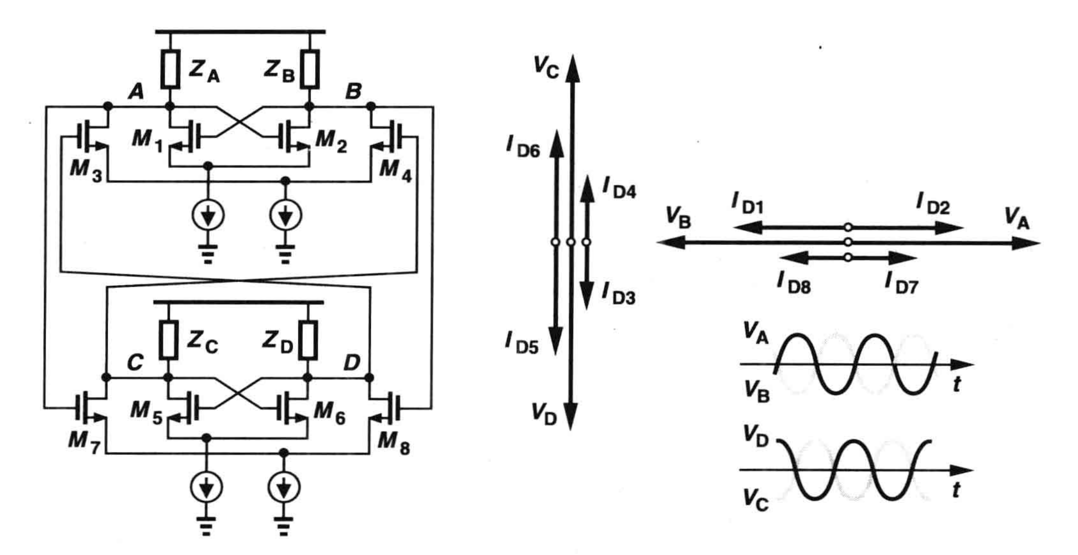


Figure 8.102 *Phasor diagrams for anti-phase coupling.*

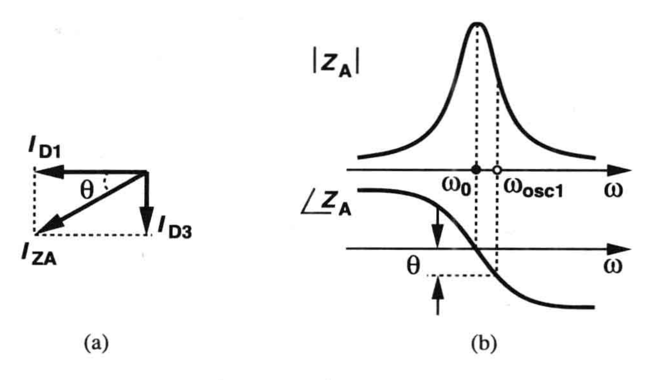


Figure 8.103 *(a) Vector summation of core and coupling currents, (b) resulting departure from resonance.*

i.e., the necessary rotation is determined by the coupling factor. If $Z_A = (L_1 s)||(C_1 s)^{-1}||R_p$, then the phase shift of Z_A can be set equal to $-\tan^{-1}(I_{D3}/I_{D1})$ so as to obtain ω_{osc1}:

$$\frac{\pi}{2} - \tan^{-1}\frac{L_1\omega_{osc1}}{R_p(1 - L_1C_1\omega_{osc1}^2)} = -\tan^{-1}\frac{I_{D3}}{I_{D1}}. \tag{8.191}$$

Since $\omega_{osc1} - \omega_0 = \Delta\omega \ll \omega_0$, the argument of $\tan^{-1}$ on the left-hand side can be simplified by writing $\omega_{osc1} \approx \omega_0$ in the numerator and $\omega_{osc1}^2 \approx \omega_0^2 + 2\Delta\omega \cdot \omega_0$ in the denominator:

$$\frac{\pi}{2} - \tan^{-1}\frac{1}{-2R_pC_1\Delta\omega} = -\tan^{-1}\frac{I_{D3}}{I_{D1}}. \tag{8.192}$$

We also recognize that $\tan^{-1} a \approx \pi/2 - 1/a$ if $a \gg 1$ and apply this approximation to the left-hand side:

$$-2R_pC_1\Delta\omega = -\tan^{-1}\frac{I_{D3}}{I_{D1}}. \tag{8.193}$$

Since $2R_pC_1\omega_0 = 2Q_{tank}$,

$$\Delta\omega = \frac{\omega_0}{2Q_{tank}} \tan^{-1} \frac{I_{D3}}{I_{D1}}. \tag{8.194}$$

Equation (8.95) yields the same result.

Example 8.45

In the phasor diagram of Fig. 8.102, we have assumed that V_C is 90° *ahead* of V_A. Is it possible for V_C to remain 90° *behind* V_A?

Solution:

Yes, it is. We construct the phasor diagram as shown in Fig. 8.104(a), noting that I_{D3} now points upward. Depicted in Fig. 8.104(b), the resultant of I_{D1} and I_{D3} must now be rotated *counterclockwise* by the tank, requiring that the oscillation frequency fall *below* ω_0 [Fig. 8.104(c)]. In this case,

$$\Delta\omega = -\frac{\omega_0}{2Q_{tank}} \tan^{-1} \frac{I_{D3}}{I_{D1}}. \tag{8.195}$$

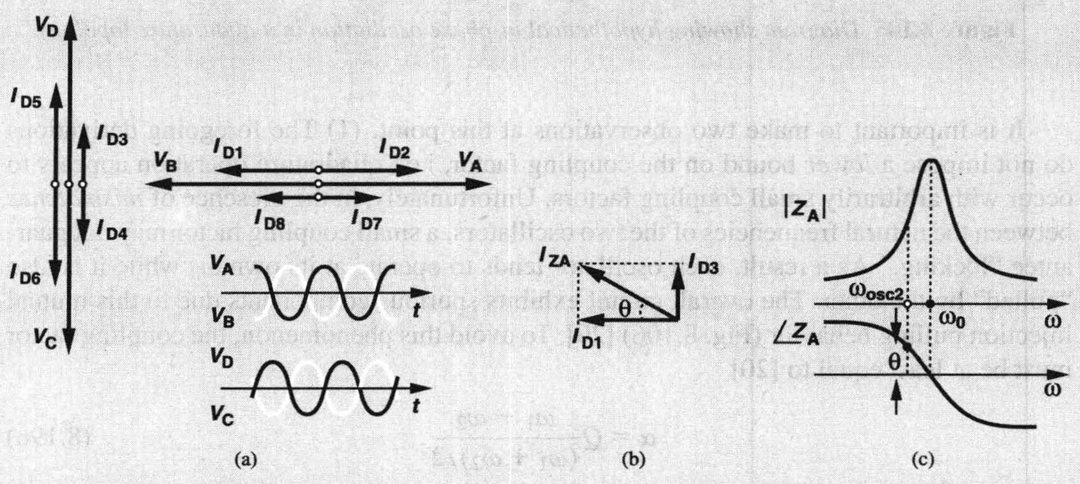


Figure 8.104 *(a) Phasors for another possible mode in quadrature oscillators, (b) vector summation of core and coupling currents, (c) resulting departure from resonance.*

The above example implies that quadrature oscillators may operate at either one of the two frequencies above and below ω_0. In fact, Eq. (8.185) also predicts the same results: since $H(j\omega_0)$ is a *complex* number, oscillation must depart from resonance, and since both $1 + j\alpha_1$ and $1 - j\alpha_1$ are acceptable, two solutions above and below resonance exist. Observed in practice as well [19], this property proves a serious drawback. In transient circuit *simulations*, the oscillator typically operates at ω_{osc2}, exhibiting little tendency to start in the higher mode even with different initial conditions. Nonetheless, it is possible

to devise a simulation that reveals the possibility of oscillation at either frequency. This is explained in Appendix A.

Example 8.46

Explain intuitively why the coupled oscillators in Fig. 8.102 cannot operate in-phase?

Solution:

If they do, then the voltage and current phasors appear as shown in Fig. 8.105. Note that I_{D3} opposes I_{D1}, whereas I_{D7} enhances I_{D5}, thereby yielding larger output swings for the bottom oscillator than for the top one. But, the symmetry of the overall circuit prohibits such an imbalance. By the same token, any phase difference other than 90° is discouraged.

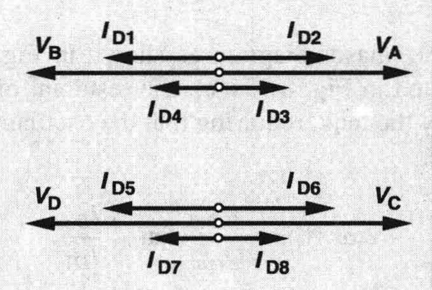


Figure 8.105 *Diagram showing hypothetical in-phase oscillation in a quadrature topology.*

It is important to make two observations at this point. (1) The foregoing derivations do not impose a *lower* bound on the coupling factor, i.e., quadrature operation appears to occur with arbitrarily small coupling factors. Unfortunately, in the presence of *mismatches* between the natural frequencies of the two oscillators, a small coupling factor may not guarantee "locking." As a result, each oscillator tends to operate at its own ω_0 while it is also "pulled" by the other. The overall circuit exhibits spurious components due to this mutual injection pulling behavior (Fig. 8.106) [20]. To avoid this phenomenon, the coupling factor must be at least equal to [20]

$$\alpha = Q\frac{\omega_1 - \omega_2}{(\omega_1 + \omega_2)/2}. \tag{8.196}$$

We typically choose an α in the range of 0.2 to 0.25. (2) As the coupling factor increases, two issues become more serious: (a) ω_{osc1} and ω_{osc2} diverge further, making it difficult to target the desired frequency range if both can occur; and (b) the phase noise of the circuit rises, with the flicker noise of the coupling transistors contributing significantly at low frequency offsets [21, 27]. This can be seen by noting that $\Delta\omega$ in Eq. (8.194) is a function of the coupling factor, I_{D3}/I_{D1}, which itself varies slowly with the flicker noise of the coupling transistors and their tail currents [21] (Problem 8.16). It follows that the choice of the coupling factor entails a trade-off between proper, spurious-free operation and phase noise. For a given power dissipation, the phase noise of quadrature oscillators is typically 3 to 5 dB higher [23] than a single oscillator.

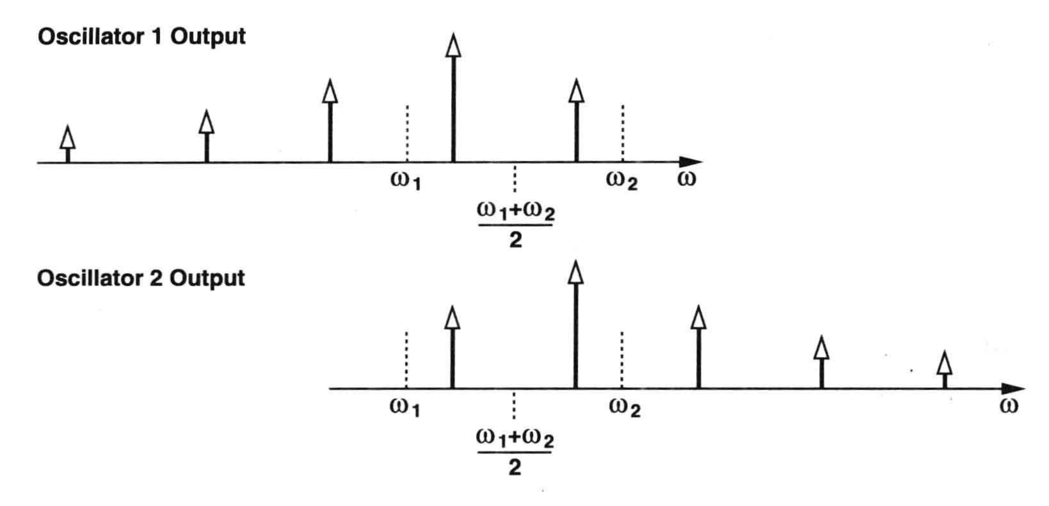


Figure 8.106 *Mutual injection pulling between two oscillators.*

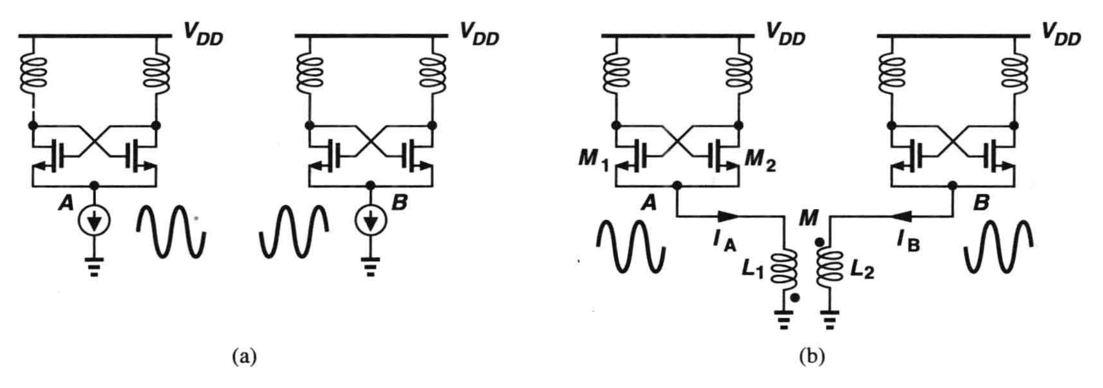


Figure 8.107 *(a) Two differential oscillators operating in-quadrature, (b) coupling through tails to ensure quadrature operation.*

The mismatches between the two oscillator cores and the coupling pairs result in phase and amplitude mismatch between the quadrature outputs. These effects are studied in [22, 24].

8.11.3 Improved Quadrature Oscillators

A number of quadrature oscillator topologies have been proposed that alleviate the trade-offs mentioned above. The principal drawback of the generic configuration studied thus far is that the coupling pairs introduce significant phase noise. We therefore postulate that, if the quadrature relationship between the two core oscillators is established by a different means, then phase noise can be reduced.

Consider the two oscillators shown in Fig. 8.107(a) and suppose they are somehow forced to operate in quadrature at a frequency of ω_{osc}. The tail nodes, A and B, thus exhibit periodic waveforms at $2\omega_{osc}$ and 180° out of phase. Conversely, if additional circuitry forces A and B to sustain a phase difference of 180°, then the two oscillators operate in

quadrature. Illustrated in Fig. 8.107(b) [25], such circuitry can be simply a 1-to-1 transformer that couples V_A to V_B and vice versa. The coupling polarity is chosen such that the transformer inverts the voltage at each node and applies it to the other.

Example 8.47

Explain what prohibits the two core oscillators in Fig. 8.107(b) from operating in-phase. Assume $L_1 = L_2$.

Solution:

Assuming a mutual coupling factor of M between L_1 and L_2, we have in the general case,

$$I_A L_1 s - I_B M s = V_A \tag{8.197}$$

$$I_B L_2 s - I_A M s = V_B. \tag{8.198}$$

If the two oscillators operate in quadrature, then $V_A = -V_B$ and $I_A = -I_B$, yielding a tail impedance of

$$\frac{V_A}{I_A} = L_1 s + M s. \tag{8.199}$$

The equivalent inductance, $L_1 + M$, is chosen such that it resonates with the tail node capacitance at $2\omega_{osc}$, thereby creating a high impedance and allowing A and B to swing freely. On the other hand, if the oscillators operate in-phase, then $V_A = V_B$ and $I_A = I_B$, giving a tail impedance of

$$\frac{V_A}{I_A} = L_1 s - M s. \tag{8.200}$$

If L_1 and L_2 are closely coupled, then $L_1 \approx M$, and Eq. (8.200) suggests that nodes A and B are almost shorted to ground for common-mode swings. The overall circuit therefore has little tendency to produce in-phase outputs.

The topology of Fig. 8.107(b) merits several remarks. First, since the coupling pairs used in the generic circuit of Fig. 8.97 are absent, the two core oscillators operate at their tanks' resonance frequency, ω_{osc}, rather than depart so as to produce additional phase shift. This important attribute means that this approach avoids the frequency ambiguity suggested by Eqs. (8.194) and (8.195). Moreover, it improves the phase noise. Second, in a manner similar to that illustrated in Fig. 8.88, the resonance of $L_1 + M$ with the tail capacitance at $2\omega_{osc}$ also improves the phase noise [25]. Third, unfortunately, the circuit requires a transformer in addition to the main tank inductors, facing a complex layout. Figure 8.108 shows a possible placement of the devices, indicating that T_1 must remain relatively far from the main inductors so as to minimize the leakage of $2\omega_{osc}$ to the two core oscillators. Such leakage distorts the duty cycle of the outputs, degrading the IP_2 of the mixers driven by such waveforms.

In the presence of mismatches between the core oscillators of Fig. 8.107(b), both the voltage swings at A and B and the mutual coupling between L_1 and L_2 must exceed a

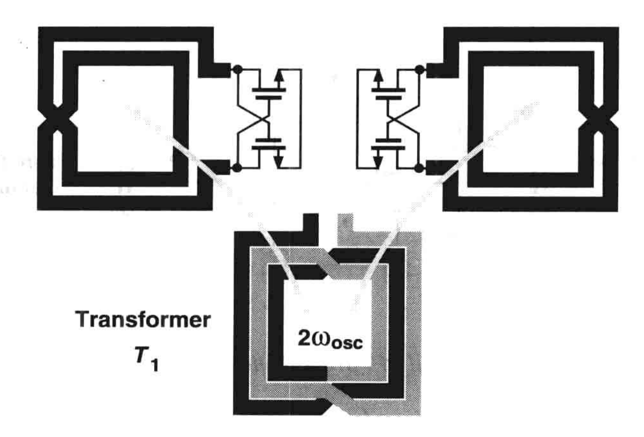


Figure 8.108 *Coupling between tail transformer and core oscillators.*

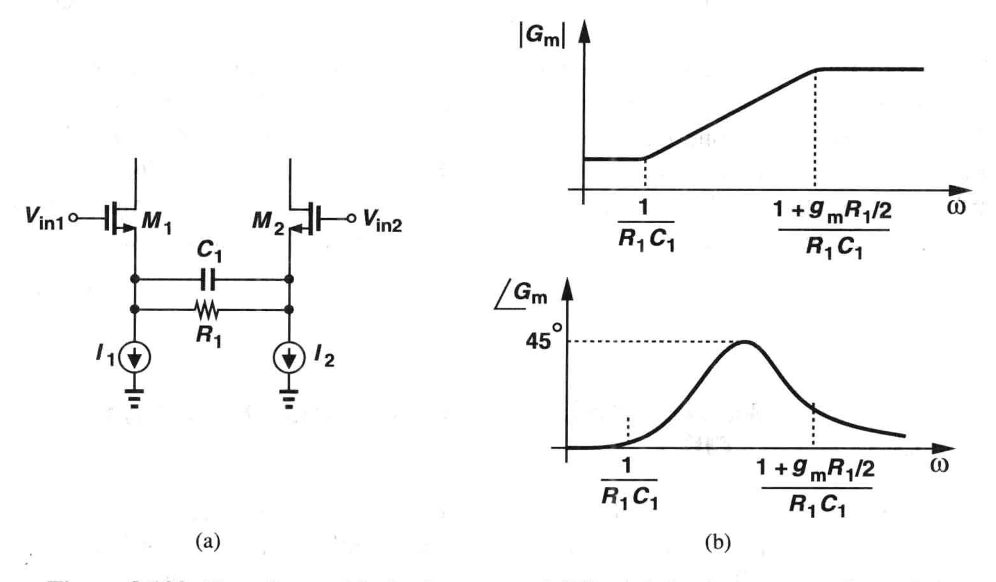


Figure 8.109 *Use of capacitively-degenerated differential pair to create phase shift.*

certain minimum to guarantee lock. Note that I_A is commutated by M_1 and M_2 (as in a mixer), experiencing a conversion gain of $2/\pi$.

The above example reveals that techniques that reduce the deviation of the oscillation frequency from the resonance frequency may also lower the flicker noise contribution of the coupling transistors. Specifically, it is possible to introduce additional phase shift in the coupling network by means of passive devices so as to drive $\Delta\omega$ in Eqs. (8.194) and (8.195) toward zero. Shown in Fig. 8.109(a) is an example where the degeneration network yields an overall transconductance of

$$G_m(s) = \frac{g_m(R_1C_1s + 1)}{R_1C_1s + 1 + g_mR_1/2}. \tag{8.201}$$

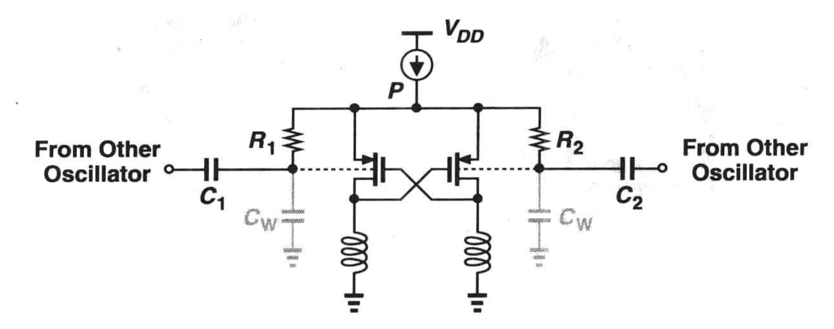


Figure 8.110 *Coupling through n-well of PMOS devices to avoid flicker noise upconversion.*

As depicted in Fig. 8.109(b), the phase reaches several tens of degrees between the zero and pole frequencies. However, the degeneration does reduce the coupling factor, requiring larger transistors and bias currents.

In order to avoid the flicker noise of the coupling devices, one can perform the coupling through the *bulk* of the main transistors. Illustrated in Fig. 8.110 [26], the idea is to apply the differential output of one oscillator to the *n*-well of the cross-coupled transistors in the other.[11] The large resistors R_1 and R_2 set the bias voltage of the *n*-wells to V_P. Also, C_1 and C_2 are small enough to create a coupling factor of about 0.25 in conjunction with the *n*-well-substrate capacitance, C_W. Note that this technique still allows two oscillation frequencies.

8.12 APPENDIX A: SIMULATION OF QUADRATURE OSCILLATORS

In order to examine the tendency of the quadrature oscillator to operate at the frequencies above and below ω_0 [Figs. 8.103(b) and 8.104(c)], we simulate the circuit as follows. First, we reconfigure the circuit so that it operates with *in-phase* coupling and hence at ω_0. This simulation provides the exact value of ω_0 in the presence of all capacitances.

Next, we apply anti-phase coupling and simulate the circuit, obtaining the exact value of ω_{osc2} (or ω_{osc1} if the oscillator prefers the higher mode). Since $\omega_0 - \omega_{osc2} \approx \omega_{osc1} - \omega_0$, we now have a relatively accurate value for ω_{osc1}.

Last, we *inject* a sinusoidal current of frequency ω_{osc1} into the oscillator, $I_{inj} = I_0 \cos \omega_{osc1} t$ (Fig. 8.111) and allow the circuit to run for a few hundred cycles. If I_0 is sufficiently large (e.g., $0.2 I_{SS}$), the circuit is likely to "lock" to ω_{osc1}. We turn *off* I_{inj} after lock is achieved and observe whether the oscillator continues to operate at ω_{osc1}. If it does, then ω_{osc1} is also a possible solution.

We should also mention that a realistic inductor model is essential to proper simulation of quadrature oscillators. Without the parallel or series resistances that model various loss mechanisms (Chapter 7), the circuit may behave strangely in simulations.

11. In [26], the transistors are NMOS devices with the assumption that their bulks can be separated.

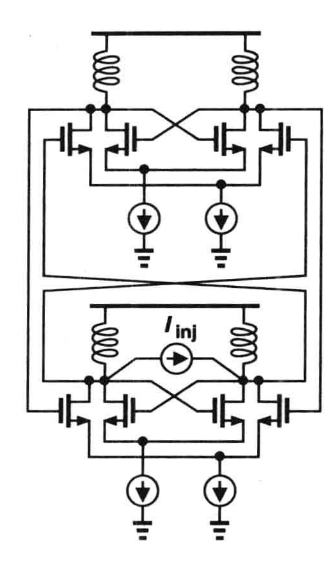


Figure 8.111 *Example of injection locking to an external current source.*

REFERENCES

[1] B. Razavi, "A 300-GHz Fundamental Oscillator in 65-nm CMOS Technology," *Symposium on VLSI Circuits Dig. Of Tech. Papers,* pp. 113–114, June 2010.

[2] R. B. Staszewski et al., "All-Digital PLL and GSM/EDGE Transmitter in 90-nm CMOS," *ISSCC Dig. Tech. Papers,* pp. 316–317, Feb. 2005.

[3] D. B. Leeson, "A Simple Model of Feedback Oscillator Noise Spectrum," *Proc. IEEE,* vol. 54, pp. 329–330, Feb. 1966.

[4] B. Razavi, "A Study of Phase Noise in CMOS Oscillators," *IEEE Journal of Solid-State Circuits,* vol. 31, pp. 331–343, March 1996.

[5] J. Craninckx and M. Steyaert, "Low-Noise Voltage-Controlled Oscillators Using Enhanced LC Tanks," *IEEE Tran. Circuits and Systems, II,* vol. 42, pp. 794–804, Dec. 1995.

[6] A. Hajimiri and T. H. Lee, "A General Theory of Phase Noise in Electrical Oscillators," *IEEE J. of Solid-State Circuits,* vol. 33, pp. 179–194, Feb. 1998.

[7] J. J. Rael and A. A. Abidi, "Physical Processes of Phase Noise in Differential LC Oscillators," *Proc. CICC,* pp. 569–572, May 2000.

[8] J. J. Rael, *Phase Noise in Oscillators,* PhD Dissertation, University of California, Los Angeles, 2007.

[9] L. W. Couch, *Digital and Analog Communication Systems,* Fourth Edition, New York: Macmillan Co., 1993.

[10] P. Andreani et al., "A Study of Phase Noise in Colpitts and LC-Tank CMOS Oscillators," *IEEE J. Solid-State Circuits,* vol. 40, pp. 1107–1118, May 2005.

[11] P. Andreani and A. Fard, "More on the 1/f Phase Noise Performance of CMOS Differential-Pair LC-Tank Oscillators," *IEEE J. Solid-State Circuits,* vol. 41, pp. 2703–2712, Dec. 2006.

[12] C. Samori et al., "Spectrum Folding and Phase Noise in LC Tuned Oscillators," *IEEE Tran. Circuits and Systems, II,* vol. 45, pp. 781–791, July 1998.

[13] S. Levantino et al., "AM-to-PM Conversion in Varactor-Tuned Oscillators," *IEEE Tran. Circuits and Systems, II,* vol. 49, pp. 509–513, July 2002.

[14] A. Bonfanti et al., "A Varactor Configuration Minimizing the Amplitude-to-Phase Noise Conversion," *IEEE Tran. Circuits and Systems, II,* vol. 53, pp. 481–488, March 2006.

[15] B. De Muer et al., "A 2-GHz Low-Phase-Noise Integrated LC-VCO Set with Flicker-Noise Upconversion Minimization," *IEEE J. of Solid-State Circuits,* vol. 35, pp. 1034–1038, July 2000.
[16] E. Hegazi, H. Sjoland, and A. A. Abidi, "A Filtering Technique to Lower LC Oscillator Phase Noise," *IEEE J. Solid-State Circuits,* vol. 36, pp. 1921–1930, Dec. 2001.
[17] A. Mazzanti and P. Andreani, "Class-C Harmonic CMOS VCOs, with a General Result on Phase Noise," *IEEE J. Solid-State Circuits,* vol. 43, no. 12, pp. 2716–2729, Dec. 2008.
[18] T. P. Liu, "A 6.5-GHz Monolithic CMOS Voltage-Controlled Oscillator," *ISSCC Dig. Tech. Papers,* pp. 404–405, Feb. 1999.
[19] S. Li, I. Kipnis, and M. Ismail, "A 10-GHz CMOS Quadrature LC-VCO for Multirate Optical Applications," *IEEE J. Solid-State Circuits,* vol. 38, pp. 1626–1634, Oct. 2003.
[20] B. Razavi, "Mutual Injection Pulling Between Oscillators," *Proc. CICC,* pp. 675–678, Sept. 2006.
[21] P. Andreani et al., "Analysis and Design of a 1.8-GHz CMOS LC Quadrature VCO," *IEEE J. Solid-State Circuits,* vol. 37, pp. 1737–1747, Dec. 2002.
[22] L. Romano et al., "Phase Noise and Accuracy in Quadrature Oscillators," *Proc. ISCAS,* pp. 161–164, May 2004.
[23] B. Razavi, "Design of Millimeter-Wave CMOS Radios: A Tutorial," *IEEE Trans. Circuits and Systems, I,* vol. 56, pp. 4–16, Jan. 2009.
[24] A Mazzanti, F. Svelto, and P. Andreani, "On the Amplitude and Phase Errors of Quadrature LC-Tank CMOS Oscillators," *IEEE J. Solid-State Circuits,* vol. 41, pp. 1305–1313, June 2006.
[25] S. Gierkink et al., "A Low-Phase-Noise 5-GHz Quadrature CMOS VCO Using Common-Mode Inductive Coupling," *Proc. ESSCIRC,* pp. 539–542, Sept. 2002.
[26] H. R. Kim et al., "A Very Low-Power Quadrature VCO with Back-Gated Coupling," *IEEE J. Solid-State Circuits,* vol. 39, pp. 952–955, June 2004.
[27] A. Mazzanti and P. Andreani, "A Time-Variant Analysis of Fundamental $1/f^3$ Phase Noise in CMOS Parallel LC-Tank Quadrature Oscillator," *IEEE Tran. Circuits and Systems, I,* vol. 56, pp. 2173–2181, Oct. 2009.
[28] B. Razavi, "Cognitive Radio Design Challenges and Techniques," *IEEE Journal of Solid-State Circuits,* vol. 45, pp. 1542–1553, Aug. 2010.

PROBLEMS

8.1. Determine the input admittance of the circuit shown in Fig. 8.4 and find its real part.

8.2. Suppose $H(s)$ in Fig. 8.6 satisfies the following conditions at a frequency ω_1: $|H(j\omega_1| = 1$ but $\angle H(j\omega_1) = 170^\circ$. Explain what happens.

8.3. Repeat the above problem if $|H(j\omega_1| < 1$ but $\angle H(j\omega_1) = 180^\circ$.

8.4. Analyze the oscillator of Fig. 8.15(b) if C_{GD} is not neglected.

8.5. Can any feedback oscillator that employs a lossy resonator be viewed as the one-port system of Fig. 8.13(c)?

8.6. Suppose the inductors in the oscillator of Fig. 8.17(a) exhibit a mismatch of ΔL. Determine the oscillation frequency by calculating the frequency at which the total phase shift around the loop reaches 360°.

8.7. Prove that the series combination of the two tanks in Fig. 8.21(a) can be replaced with one tank as shown in Fig. 8.21(b).

8.8. Compute the tuning range in Example 8.18 if C_b is placed at nodes P and Q in Fig. 8.32(a).

8.9. Why do the PMOS devices in Fig. 8.36 carry a current of I_{SS}?

8.10. For a CS stage loaded by a second-order parallel RLC tank, prove that $R_p/(L\omega_0) = (\omega_0/2)d\phi/d\omega(=Q)$.

8.11. Prove that the noise shaping in the system shown in Fig. 8.59 is given by Eq. (8.108).

8.12. Assuming $x(t) = A\cos\omega_0 t + n(t) = A\cos\omega_0 t + n_I(t)\cos\omega_0 t - n_Q(t)\sin\omega_0 t$, show that the power carried by the AM sidebands is equal to that carried by the PM sidebands and equal to half of the power of $n(t)$.

8.13. Suppose the VCO of Fig. 8.25(a) employs varactors whose capacitance is given by $C_{var} = C_0(1 + \alpha_1 V + \alpha_2 V^2)$, where V denotes the gate-source voltage. Assume complete current steering and a low-frequency noise current of $I_n = I_m \cos\omega_m t$ in I_{SS}.

(a) Determine the AM noise resulting from I_n.

(b) Determine how the average value of the varactor capacitance varies with I_n.

(c) Compute the phase modulation as a result of the tank resonance frequency modulation.

8.14. Prove that the peak drain voltage swing in Fig. 8.86(b) is no more than roughly $V_{DD} - 2(V_{GS} - V_{TH})$. To approach this value, the capacitive attenuation must minimize the gate voltage swing.

8.15. Assuming $Z_T = (L_1 s)||(C_1 s)^{-1}||R_p$ in Fig. 8.100, determine the startup condition.

8.16. For small-signal operation, Eq. (8.194) can be written as

$$\Delta\omega = \frac{\omega_0}{2Q_{tank}} \tan^{-1}\frac{g_{m3}}{g_{m1}}. \tag{8.202}$$

Now suppose the tail current of the coupling transistors, I_{T1} contains a flicker noise component, $I_n \ll I_{T1}$. Writing $g_{m3} = \sqrt{2\mu_n C_{ox}(W/L)_3(I_{T1} + I_n)/2}$, express $\Delta\omega$ as a linear function of I_n and obtain the corresponding "gain," $K_{VCO} = \partial(\Delta\omega)/\partial I_n$.

8.17. In the VCO circuit shown in Fig. 8.112, the voltage dependence of each varactor can be expressed as $C_{var} = C_0(1 + \alpha_1 V_{var})$, where V_{var} denotes the average voltage across the varactor. Use the narrowband FM approximation in this problem. Also, neglect all other capacitances and assume the circuit oscillates at a frequency of ω_0 for the given value of V_{cont}. The dc drop across the inductors is negligible.

(a) Compute the "gain" from I_{SS} to the output frequency, ω_{out}. That is, assume I_{SS} changes by a small value and calculate the voltage change across the varactors and hence the change in the output frequency.

(b) Assume I_{SS} has a noise component that can be expressed as $I_n \cos\omega_n t$. Using the result found in (a), determine the frequency and relative magnitude of the resulting output sidebands of the oscillator.

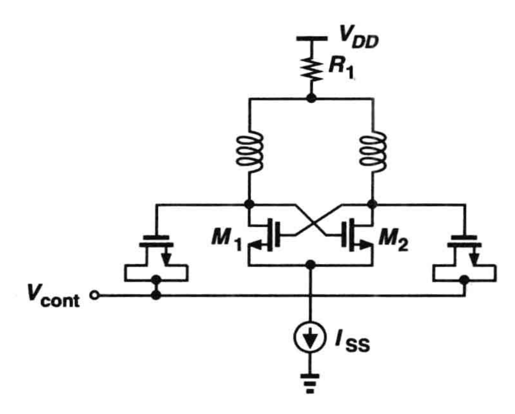


Figure 8.112 *VCO with level-shift resistor.*

8.18. The circuit shown in Fig. 8.113 is a simplified model of a "dual-mode" oscillator [28]. The voltage-dependent current source models a transistor. The circuit oscillates if Z_{in} goes to infinity for $s = j\omega$.

(a) Determine the input impedance Z_{in}.

(b) Set the denominator of Z_{in} to zero for $s = j\omega$ and obtain the startup condition and the two oscillation frequencies.

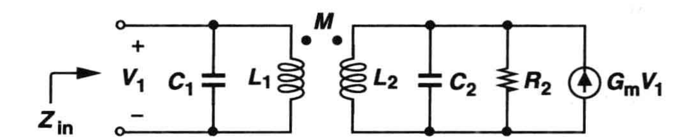


Figure 8.113 *Simplified model of a dual-mode oscillator.*

CHAPTER



PHASE-LOCKED LOOPS

Most synthesizers employ "phase-locking" to achieve high frequency accuracy. We therefore dedicate this chapter to a study of PLLs. While a detailed treatment of PLLs would consume an entire book, our objective here is to develop enough foundation to allow the analysis and design of RF synthesizers. The outline of the chapter is shown below. The reader is encouraged to review the mathematical model of VCOs described in Chapter 8.

Type–I PLLs	Type–II PLLs	PLL Nonidealities
▪ VCO Phase Alignment	▪ Phase/Frequency Detectors	▪ PFD/CP Nonidealities
▪ Dynamics of Type–I PLLs	▪ Charge Pumps	▪ Circuit Techniques
▪ Frequency Multiplication	▪ Charge–Pump PLLs	▪ VCO Phase Noise
▪ Drawbacks of Type–I PLL	▪ Transient Response	▪ Reference Phase Noise

9.1 BASIC CONCEPTS

In its simplest form, a PLL is a negative feedback loop consisting of a VCO and a "phase detector" (PD). We therefore first define what a PD is and subsequently construct the loop.

9.1.1 Phase Detector

A PD is a circuit that senses two periodic inputs and produces an output whose average value is proportional to the difference between the *phases* of the inputs. Shown in Fig. 9.1, the input/output characteristic of the PD is ideally a straight line, with a slope called the "gain" and denoted by K_{PD}. For an output voltage quantity, K_{PD} is expressed in V/rad. In practice, the characteristic may not be linear or even monotonic.

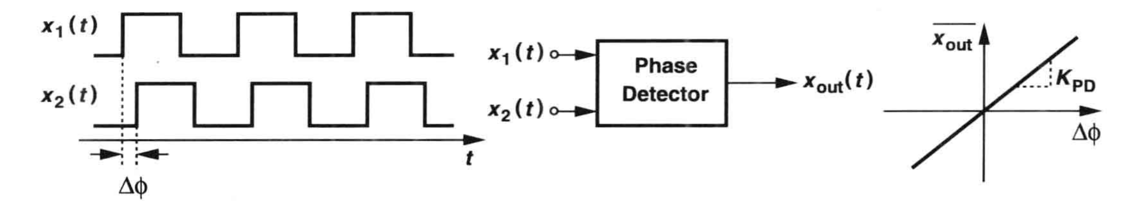


Figure 9.1 *Phase detector and its input/output characteristic.*

Example 9.1

Must the two periodic inputs to a PD have equal frequencies?

Solution:

They need not, but with unequal frequencies, the phase difference between the inputs varies with time. Figure 9.2 depicts an example, where the input with a higher frequency, $x_2(t)$, accumulates phase faster than $x_1(t)$, thereby changing the phase difference, $\Delta\phi$. The PD output pulsewidth continues to increase until $\Delta\phi$ crosses 180°, after which it decreases toward zero. That is, the output waveform displays a "beat" behavior having a frequency equal to the difference between the input frequencies. Also, note that the average phase difference is zero, and so is the average output.

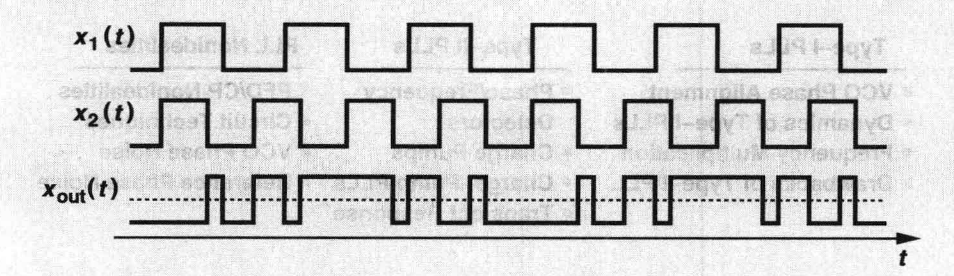


Figure 9.2 *Beating of two inputs with unequal frequencies.*

How is the phase detector implemented? We seek a circuit whose average output is proportional to the input phase difference. For example, an exclusive-OR (XOR) gate can serve this purpose. As shown in Fig. 9.3, the XOR gate generates pulses whose width is equal to $\Delta\phi$. In this case, the circuit produces pulses at both the rising edge and the falling edge of the inputs.

Example 9.2

Plot the input/output characteristic of the XOR PD for two cases: (a) the circuit has a single-ended output that swings between 0 and V_{DD}, (b) the circuit has a differential output that swings between $-V_0$ and $+V_0$.

Example 9.2 (*Continued*)

Solution:

(a) Assigning a swing of V_{DD} to the output pulses shown in Fig. 9.3, we observe that the output average begins from zero for $\Delta\phi = 0$ and rises toward V_{DD} as $\Delta\phi$ approaches 180° (because the overlap between the input pulses approaches zero). As $\Delta\phi$ exceeds 180°, the output average falls, reaching zero at $\Delta\phi = 360^\circ$. Figure 9.4(a) depicts the behavior, revealing a periodic, nonmonotonic characteristic.

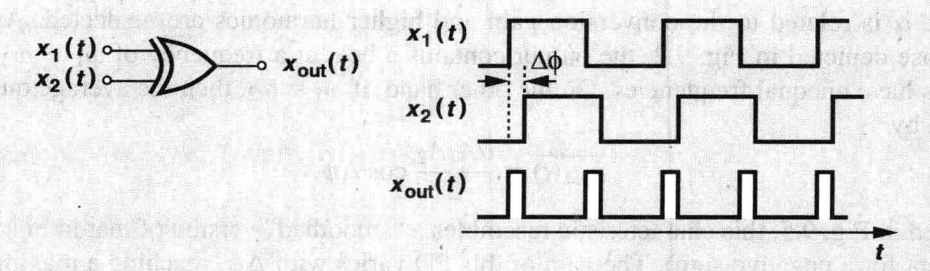


Figure 9.3 *XOR gate as a PD.*

(b) Plotted in Fig. 9.4(b) for a small phase difference, the output exhibits narrow pulses above $-V_0$ and hence an average nearly equal to $-V_0$. As $\Delta\phi$ increases, the output spends more time at $+V_0$, displaying an average of *zero* for $\Delta\phi = 90^\circ$. The average continues to increase as $\Delta\phi$ increases and reaches a maximum of $+V_0$ at $\Delta\phi = 180^\circ$. As shown in Fig. 9.4(c), the average falls thereafter, crossing zero at $\Delta\phi = 270^\circ$ and reaching $-V_0$ at 360°.

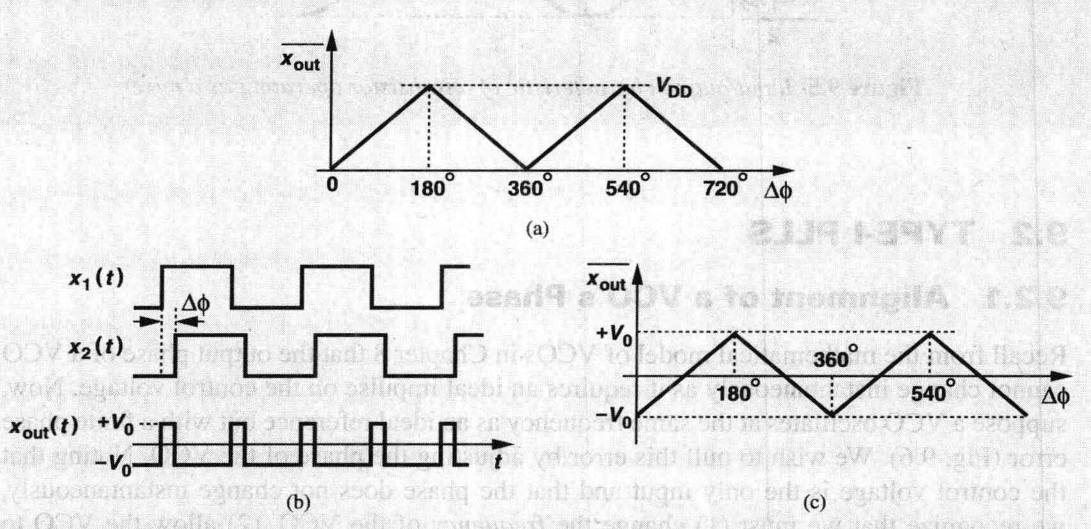


Figure 9.4 *(a) Input/output characteristic of XOR PD with output swinging between 0 and V_{DD}, (b) input and output waveforms swinging between $-V_0$ and $+V_0$, (c) characteristic corresponding to waveforms of part (b).*

Example 9.3

A single MOS switch can operate as a "poor man's phase detector." Explain how.

Solution:

Our study of mixers in Chapter 6 indicates that a MOS switch can serve as a return-to-zero or a sampling mixer. For two signals $x_1(t) = A_1 \cos \omega_1 t$ and $x_2(t) = A_2 \cos(\omega_2 t + \phi)$, the mixer generates

$$x_{out}(t) = \alpha A_1 \cos \omega_1 t \cdot A_2 \cos(\omega_2 t + \Delta\phi), \tag{9.1}$$

where α is related to the conversion gain and higher harmonics are neglected. As with the case depicted in Fig. 9.2, the output contains a beat at a frequency of $\omega_1 - \omega_2$ if the inputs have unequal frequencies. On the other hand, if $\omega_1 = \omega_2$, then the average output is given by

$$\overline{x_{out}(t)} = \frac{\alpha A_1 A_2}{2} \cos \Delta\phi. \tag{9.2}$$

Plotted in Fig. 9.5, this characteristic resembles a "smoothed" version of that in Fig. 9.4(c) (except for a negative sign). The gain of this PD varies with $\Delta\phi$, reaching a maximum of $\pm\alpha A_1 A_2/2$ at odd multiples of $\pi/2$.

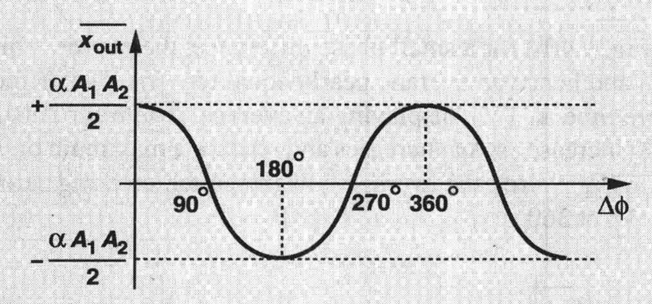


Figure 9.5 *Input/output characteristic of a transistor operating as a mixer.*

9.2 TYPE-I PLLS

9.2.1 Alignment of a VCO's Phase

Recall from the mathematical model of VCOs in Chapter 8 that the output phase of a VCO cannot change instantaneously as it requires an ideal impulse on the control voltage. Now, suppose a VCO oscillates at the same frequency as an ideal reference but with a finite phase error (Fig. 9.6). We wish to null this error by adjusting the phase of the VCO. Noting that the control voltage is the only input and that the phase does not change instantaneously, we recognize that we must (1) change the *frequency* of the VCO, (2) allow the VCO to accumulate phase faster (or more slowly) than the reference so that the phase error vanishes, and (3) change the frequency back to its initial value. As shown in Fig. 9.6, V_{cont} is stepped at $t = t_0$ and remains at the new value until $t = t_1$, when the phase error goes to zero. Thereafter, the two signals have equal frequencies and a zero phase difference.

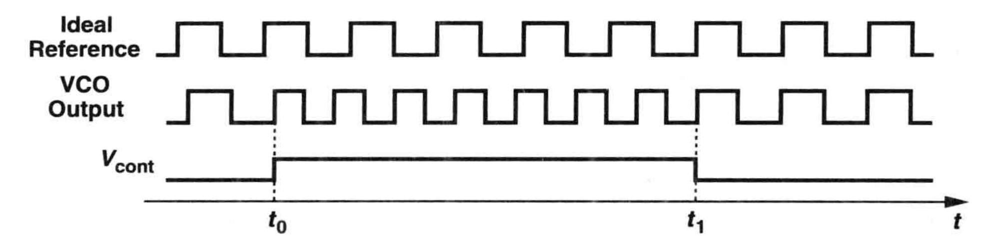


Figure 9.6 *Alignment of VCO output phase by changing its frequency.*

How do we determine the time at which the phase error in Fig. 9.6 reaches zero? A phase detector comparing the VCO phase and the reference phase can serve this purpose, yielding the negative feedback loop shown in Fig. 9.7(a). If the "loop gain" is sufficiently high, the circuit minimizes the input error. Note that the loop only "understands" phase quantities (rather than voltage or current quantities) because the input "subtractor" (the PD) operates with phases.

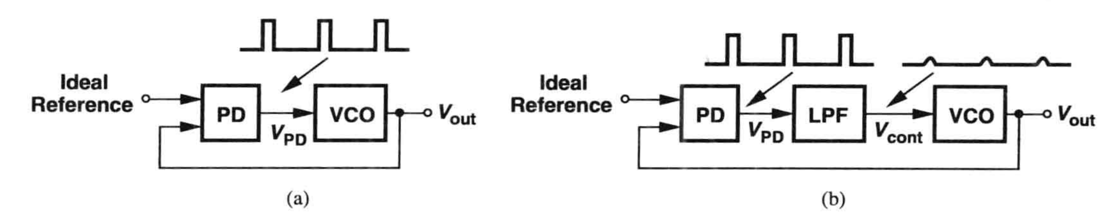


Figure 9.7 *(a) Simple PLL, (b) addition of low-pass filter to remove high-frequency components generated by PD.*

The circuit of Fig. 9.7(a) suffers from a critical issue. The phase detector produces repetitive pulses at its output, modulating the VCO frequency and generating large side-bands. We therefore interpose a low-pass filter (called the "loop filter") between the PD and the VCO so as to suppress these pulses [Fig. 9.7(b)].

Example 9.4

A student reasons that the negative feedback loop must force the phase error to *zero*, in which case the PD generates *no* pulses and the VCO is not disturbed. Thus, a low-pass filter is not necessary.

Solution:

As explained later, this feedback system suffers from a finite loop gain, exhibiting a finite phase error in the steady state. Even PLLs having an infinite loop gain contain nonidealities that disturb V_{cont}.

9.2.2 Simple PLL

We call the circuit of Fig. 9.7(b) a phase-locked loop and will study its behavior in great detail. But it is helpful to decipher the expressions "phase-locked" or "phase locking."

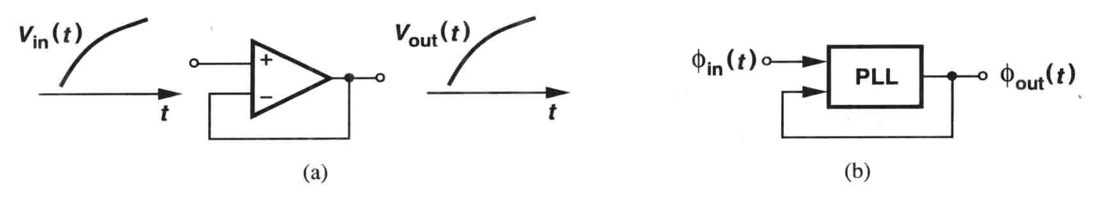


Figure 9.8 *(a) Unity-gain voltage buffer with its output tracking its input, (b) PLL with its output tracking its input.*

First, consider the more familiar voltage-domain circuit shown in Fig. 9.8(a). If the open-loop gain of the unity-gain buffer is relatively large, then the output voltage "tracks" the input voltage. Similarly, the PLL of Fig. 9.8(b) ensures that $\phi_{out}(t)$ tracks $\phi_{in}(t)$. We say the loop is "locked" if $\phi_{out}(t) - \phi_{in}(t)$ is *constant* (not necessarily zero) with time. We also say the output phase is "locked" to the input phase to emphasize the tracking property.

An important and unique consequence of phase locking is that the input and output frequencies of the PLL are *exactly* equal. This can be seen by writing

$$\phi_{out}(t) - \phi_{in}(t) = \text{constant}, \tag{9.3}$$

and hence

$$\frac{d\phi_{out}}{dt} = \frac{d\phi_{in}}{dt}. \tag{9.4}$$

This attribute proves critical to the operation of phase-locked systems, including RF synthesizers.

Example 9.5

A student argues that the input and output frequencies are exactly equal even if the phase detector in Fig. 9.7(b) is replaced with a "frequency detector" (FD), i.e., a circuit that generates a dc value in proportion to the input frequency difference. Explain the flaw in this argument.

Solution:

Figure 9.9 depicts the student's idea. We may call this a "frequency-locked loop" (FLL). The negative-feedback loop attempts to minimize the error between f_{in} and f_{out}. But, does this error fall to zero? This circuit is analogous to the unity-gain buffer of Fig. 9.8(a), whose input and output may not be exactly equal due to the *finite gain* and *offset* of the op amp. The FLL may also suffer from a finite error if its loop gain is finite or if the frequency detector exhibits offsets. [Similarly, the PLL of Fig. 9.7(b) may not yield $\phi_{out}(t) = \phi_{in}(t)$, but, as a by-product of phase locking, it guarantees that $f_{out} = f_{in}$.]

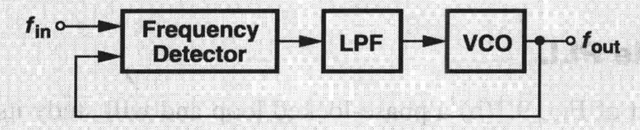


Figure 9.9 *Frequency-locked loop.*

Can two periodic waveforms have a constant phase difference but different frequencies? If we define the phase difference as the time elapsed between consecutive zero crossings, we observe that this is not possible. That is, if the phases are "locked," then the frequencies are naturally equal.

9.2.3 Analysis of Simple PLL

Figure 9.10(a) shows a PLL implementation using an XOR gate and a top-biased LC VCO (Chapter 8). The low-pass filter is realized by means of R_1 and C_1. If the loop is locked, the input and output frequencies are equal, the PD generates repetitive pulses, the loop filter extracts the average level, and the VCO senses this level so as to operate at the required frequency. Note that the signal of interest changes dimension as we "walk" around the loop: the PD input is a phase quantity, the PD output and the LPF output are voltage quantities, and the VCO output is a phase quantity. By contrast, the unit-gain buffer of Fig. 9.8(a) contains signals in only the voltage and current domains.

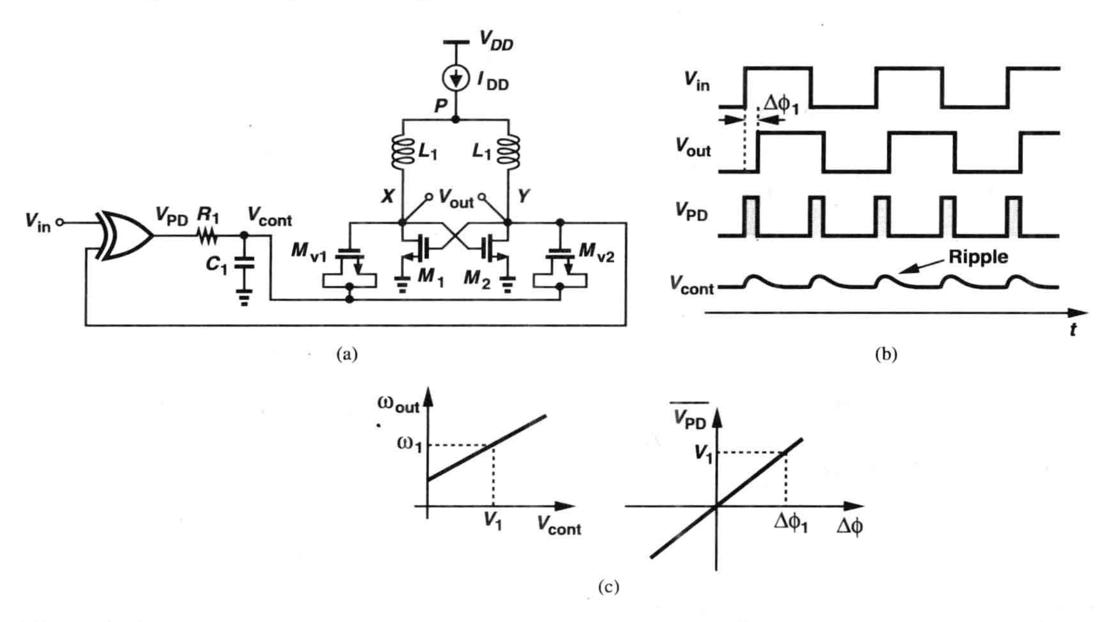


Figure 9.10 *(a) PLL implementation example, (b) waveforms at different nodes, (c) VCO and PD input/output characteristics showing the system solution.*

Following our above study, we may have many questions in regards to PLLs: (1) how does a PLL reach the locked condition? (2) does a PLL always lock? (3) how do we compute the voltages and phases around the loop in the locked condition? (4) how does a PLL respond to a change at its input? In this section, we address some of these questions.

We begin our analysis by examining the signals at various nodes in the circuit of Fig. 9.10(a). Figure 9.10(b) shows the waveforms, assuming the loop is locked. The input and output have equal frequencies but a finite phase difference, $\Delta\phi_1$, and the PD generates pulses whose width is equal to $\Delta\phi_1$. These pulses are low-pass filtered to produce the dc voltage that enables the VCO to operate at a frequency equal to the input frequency, ω_1. The residual disturbance on the control line is called the "ripple." A lower LPF corner frequency further attenuates the ripple, but at the cost of other performance parameters. We return to this point later.

With the VCO and PD characteristics known, it is possible to compute the control voltage of the VCO and the phase error. As illustrated in Fig. 9.10(c), the VCO operates at ω_1 if $V_{cont} = V_1$, and the PD generates a dc value equal to V_1 if $\Delta\phi = \Delta\phi_1$. This quantity is called the "static phase error."

Example 9.6

If the input frequency changes by $\Delta\omega$, how much is the change in the phase error? Assume the loop remains locked.

Solution:

Depicted in Fig. 9.11, such a change requires that V_{cont} change by $\Delta\omega/K_{VCO}$. This in turn necessitates a phase error change of

$$\Delta\phi_2 - \Delta\phi_1 = \frac{\Delta\omega}{K_{PD}K_{VCO}}. \tag{9.5}$$

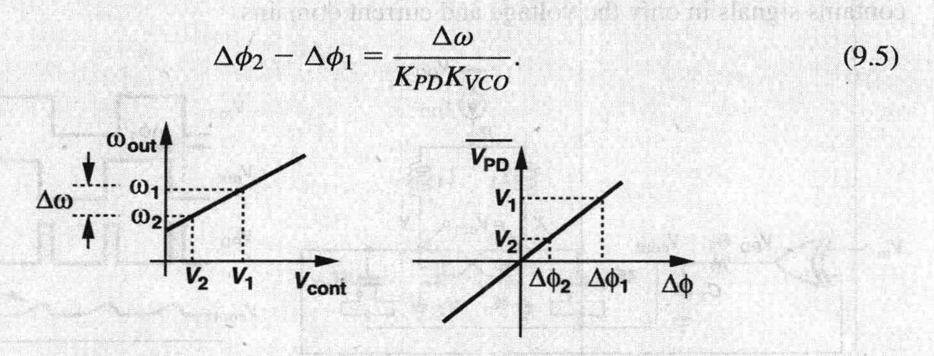


Figure 9.11 *Effect of input frequency change on phase error.*

The key observation here is that the phase error varies with the frequency. To minimize this variation, $K_{PD}K_{VCO}$ must be maximized. This quantity is sometimes called the "loop gain" even though it is not dimensionless.

Let us now study, qualitatively, the response of a PLL that is locked for $t < t_0$ and experiences a small, positive frequency step, $\Delta\omega$, at the input at $t = t_0$ (Fig. 9.12). We expect that the loop reaches the final values stipulated in Example 9.6, but we wish to examine the transient behavior. Since the input frequency, ω_{in}, is momentarily greater than the output frequency, ω_{out}, V_{in} accumulates phase faster, i.e., the phase error begins to grow. Thus, the PD generates increasingly wider pulses, raising the dc level at the output of the LPF and hence the VCO frequency. As the difference between ω_{out} and ω_{in} diminishes, so does the width of the PD output pulses, eventually settling to a value equal to $\Delta\omega/(K_{PD}K_{VCO})$ above its initial value. Also, the control voltage increases by $\Delta\omega/K_{VCO}$.

The foregoing study leads to two important points. First, among various nodes in a PLL, the control voltage provides the most straightforward representation of the transient response. By contrast, the VCO or PD outputs do not readily reveal the loop's settling behavior. Second, the loop locks only after *two* conditions are satisfied: (1) ω_{out} becomes equal to ω_{in}, and (2) the difference between ϕ_{in} and ϕ_{out} settles to its proper value [1]. For example, the plots in Fig. 9.11 reveal that an input frequency change to $\omega_1 + \Delta\omega$ demands an output frequency change to $\omega_1 + \Delta\omega$ *and* a phase error change to $\Delta\phi_2$. We also observe

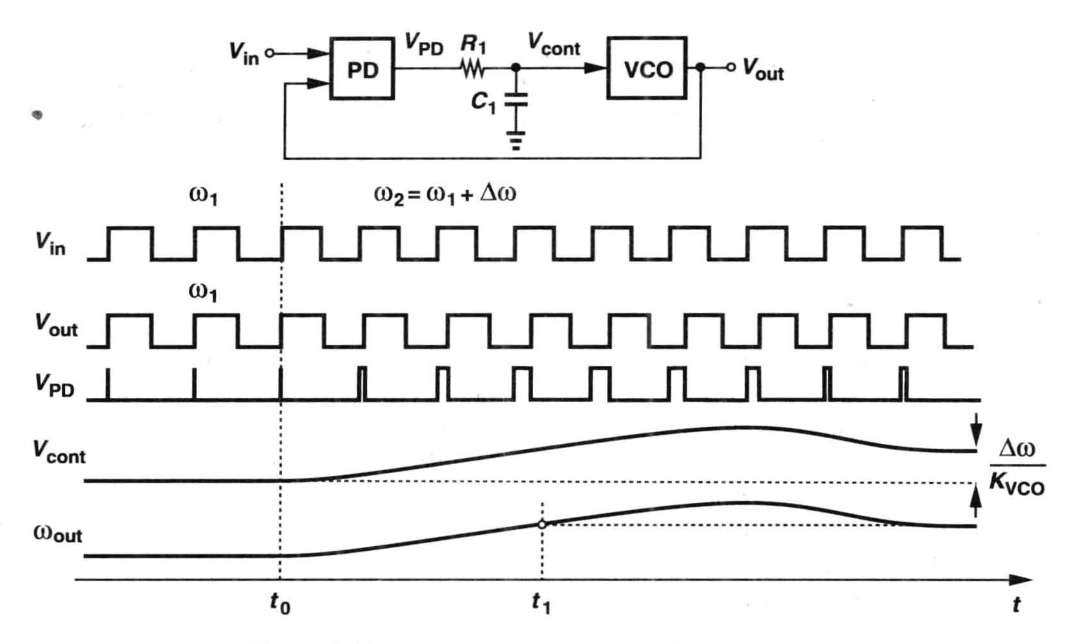


Figure 9.12 *Response of PLL to input frequency step.*

from Fig. 9.12 that V_{cont} becomes equal to its final value at $t = t_1$ (i.e., $\omega_{out} = \omega_{in}$ at this moment), but the loop continues the transient because the static phase error has not reached its proper value. In other words, both "frequency acquisition" and "phase acquisition" must be completed.

Example 9.7

An FSK waveform is applied to a PLL. Sketch the control voltage as a function of time.

Solution:

The input frequency toggles between two values and so does the output frequency. The control voltage must also toggle between two values. The control voltage waveform therefore appears as shown in Fig. 9.13, providing the original bit stream. That is, a PLL can serve as an FSK (and, more generally, FM) demodulator if V_{cont} is considered the output.

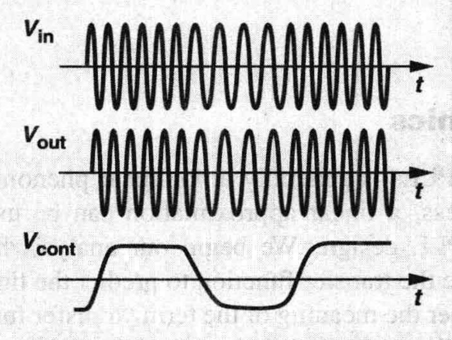


Figure 9.13 *PLL as FSK demodulator.*

Example 9.8

Having carefully followed our studies thus far, a student reasons that, except for the FSK demodulator application, a PLL is no better than a *wire* since it attempts to make the input and output frequencies and phases equal! What is the flaw in the student's argument?

Solution:

We will better appreciate the role of phase locking later in this chapter. Nonetheless, we can observe that the *dynamics* of the loop can yield interesting and useful properties. Suppose in Example 9.7, the input frequency toggles at a relatively high rate, leaving little time for the PLL to "keep up." As illustrated in Fig. 9.14, at each input frequency jump, the control voltage begins to change in the opposite direction but does not have enough time to settle. In other words, the output frequency excursions are *smaller* than the input frequency jumps. The loop thus performs *low-pass filtering* on the input frequency variations—just as the unity-gain buffer of Fig. 9.8(a) performs low-pass filtering on the input *voltage* variations if the op amp has a limited bandwidth. In fact, many applications incorporate PLLs to reduce the frequency or phase noise of a signal by means of this low-pass filtering property.

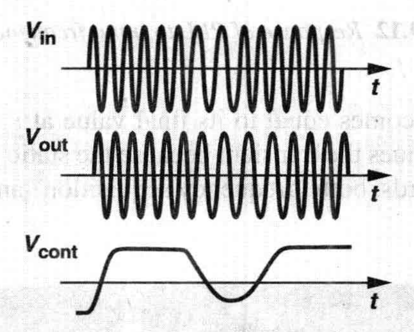


Figure 9.14 *Distortion of demodulated FSK signal due to limited PLL bandwidth.*

If the input/output phase error of a PLL varies with time, we say the loop is "unlocked," an undesirable state because the output does not track the input. For example, if at the startup, the VCO frequency is far from the input frequency, the loop may never lock. While the behavior of a PLL in the unlocked state is not important per se, whether and how it acquires lock are both critical issues. In our development of PLLs in this section, we devise a method to guarantee lock.

9.2.4 Loop Dynamics

The transient response of PLLs is generally a nonlinear phenomenon that cannot be formulated easily. Nevertheless, a linear approximation can be used to gain intuition and understand trade-offs in PLL design. We begin our analysis by obtaining the transfer function. Next, we examine the transfer function to predict the time-domain behavior.

It is instructive to ponder the meaning of the term "transfer function" in a phase-locked system. In the more familiar voltage-domain circuits, such as the unity-gain buffer of

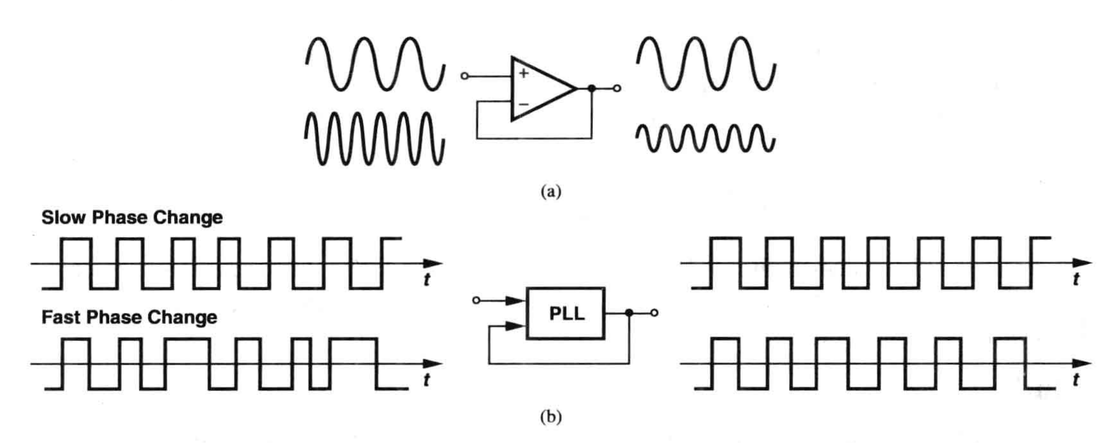


Figure 9.15 *(a) Response of unity-gain voltage buffer to low or high frequencies, (b) response of PLL to slow or fast input phase changes.*

Fig. 9.15(a), the transfer function signifies how a sinusoidal input *voltage* propagates to the output.[1] For example, a slow input sinusoid experiences little attenuation, whereas a fast sinusoid emerges with a small *voltage* amplitude. How do we extend these concepts to the *phase* domain? The transfer function of a PLL must reveal how a slow or a fast change in the input (excess) *phase* propagates to the output. Figure 9.15(b) illustrates examples of slow and fast phase change. From Example 9.7, we predict that the PLL's low-pass behavior "attenuates" the phase excursions if the input phase varies fast. That is, the output phase tracks the input phase closely only for slow phase variations.

Let us now construct a "phase-domain model" for the PLL. The phase detector simply *subtracts* the output phase from the input phase and scales the result by a factor of K_{PD} so as to generate an average *voltage*. As shown in Fig. 9.16, this voltage is applied to the low-pass filter and subsequently to the VCO. Since the phase detector only senses the output phase, the VCO must be modeled as a circuit with a voltage input and a phase output. From the model developed in Chapter 8, the VCO transfer function is expressed as K_{VCO}/s. The open-loop transfer function of the PLL is therefore given by $[K_{PD}/(R_1C_1s + 1)](K_{VCO}/s)$, yielding an overall closed-loop transfer function of

$$H(s) = \frac{\phi_{out}}{\phi_{in}}(s) = \frac{K_{PD}K_{VCO}}{R_1C_1s^2 + s + K_{PD}K_{VCO}}. \tag{9.6}$$

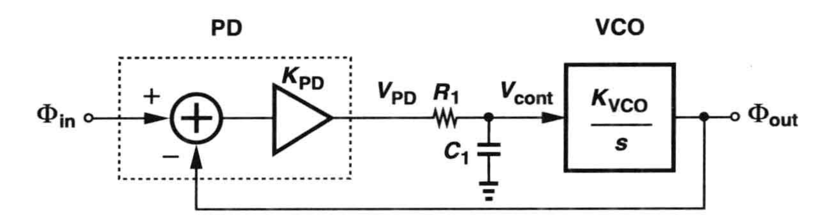


Figure 9.16 *Phase-domain model of type-I PLL.*

1. Of course, the transfer function represents the behavior for nonsinusoidal inputs as well.

Since the open-loop transfer function contains one pole at the origin (due to the VCO) (i.e., one ideal integrator), this system is called a "type-I PLL." As expected, for slow input phase variations ($s \approx 0$), $H(s) \approx 1$, i.e., the output phase tracks the input phase.

Example 9.9

The analysis illustrated in Fig. 9.10 suggests that the loop locks with a *finite* phase error, whereas Eq. (9.6) implies that $\phi_{out} = \phi_{in}$ for very slow phase variations. Are these two observations consistent?

Solution:

Yes, they are. As with any transfer function, Eq. (9.6) deals with *changes* in the input and the output rather than with their total values. In other words, (9.6) merely indicates that a phase step of $\Delta\phi$ at the input eventually appears as a phase change of $\Delta\phi$ at the output, but it does not provide the static phase offset.

The second-order transfer function given by Eq. (9.6) can have an overdamped, critically-damped, or underdamped behavior. To derive the corresponding conditions, we express the denominator in the familiar control theory form, $s^2 + 2\zeta\omega_n s + \omega_n^2$, where ζ is the "damping factor" and ω_n the "natural frequency." Thus,

$$H(s) = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}, \tag{9.7}$$

where

$$\zeta = \frac{1}{2}\sqrt{\frac{\omega_{LPF}}{K_{PD}K_{VCO}}} \tag{9.8}$$

$$\omega_n = \sqrt{K_{PD}K_{VCO}\omega_{LPF}}, \tag{9.9}$$

and $\omega_{LPF} = 1/(R_1C_1)$. The damping factor is typically chosen to be $\sqrt{2}/2$ or larger so as to provide a well-behaved (critically damped or overdamped) response.

Example 9.10

Using Bode plots of the open-loop system, explain why ζ is inversely proportional to K_{VCO}.

Solution:

Figure 9.17 shows the behavior of the open-loop transfer function, H_{open}, for two different values of K_{VCO}. As K_{VCO} increases, the unity-gain frequency rises, thus reducing the phase

Example 9.10 *(Continued)*

margin (PM). This trend is similar to that in more familiar voltage (or current) feedback circuits, where a higher loop gain leads to less stability.

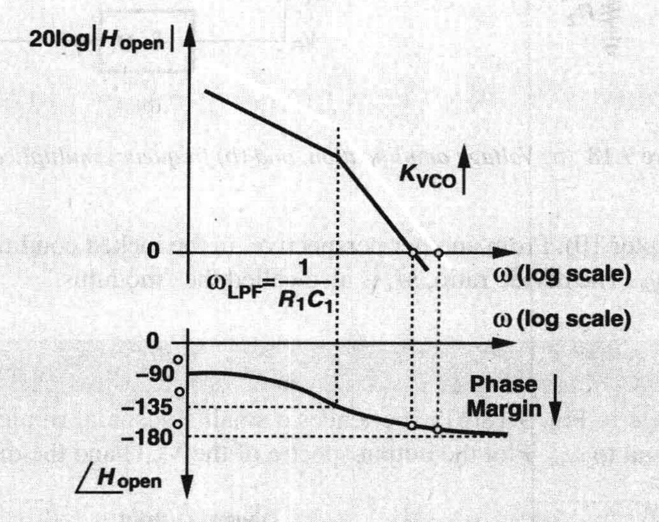


Figure 9.17 *Bode plots of type-I PLL showing the effect of higher K_{VCO}.*

Since phase and frequency are related by a linear, time-invariant operation, Eq. (9.6) also applies to frequency quantities. For example, if the input frequency varies slowly, the output frequency tracks it closely.

Example 9.11

How do we ensure the feedback in Fig. 9.10 is negative?

Solution:

As seen in Fig. 9.4(a), the phase detector provides both negative and positive gains. Thus, the loop automatically locks with negative feedback.

9.2.5 Frequency Multiplication

An extremely useful property of PLLs is frequency multiplication, i.e., the generation of an output frequency that is a multiple of the input frequency. How can a PLL "amplify" a frequency? We revisit the more familiar voltage buffer of Fig. 9.8(a) and note that it can provide amplification if its output is *divided* (attenuated) before returning to the input [Fig. 9.18(a)]. Similarly, the output frequency of a PLL can be divided and then fed back [Fig. 9.18(b)]. The $\div M$ circuit is a counter that generates one output pulse for every M

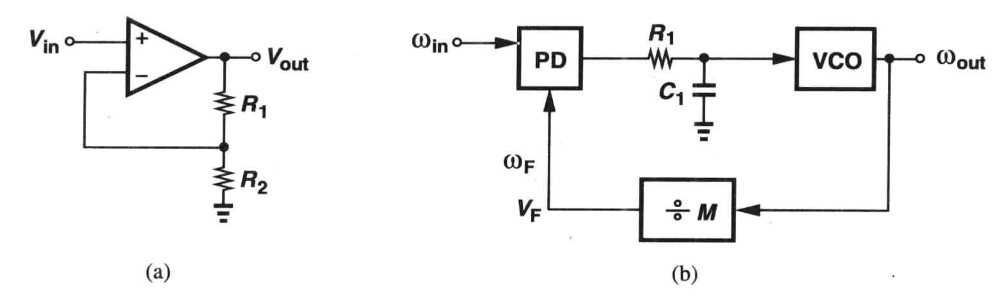


Figure 9.18 *(a) Voltage amplification, and (b) frequency multiplication.*

input pulses (Chapter 10). From another perspective, in the locked condition, $\omega_F = \omega_{in}$ and hence $\omega_{out} = M\omega_{in}$. The divide ratio, M, is also called the "modulus."

Example 9.12

The control voltage in Fig. 9.18(b) experiences a small sinusoidal ripple of amplitude V_m at a frequency equal to ω_{in}. Plot the output spectra of the VCO and the divider.

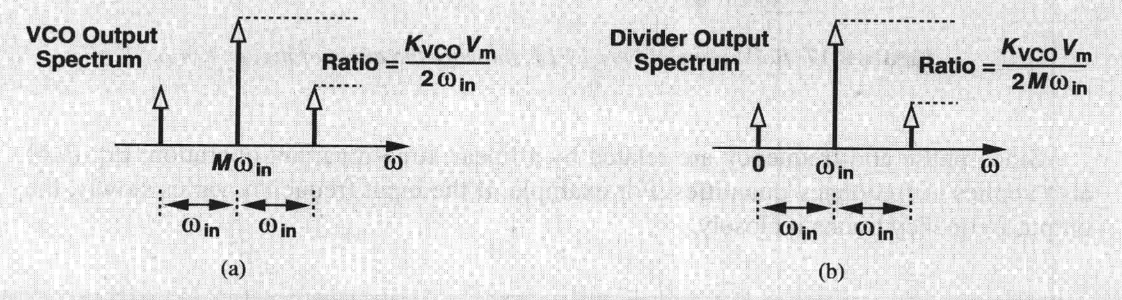


Figure 9.19 *Spectra at (a) VCO output, and (b) divider output.*

Solution:

From the narrowband FM approximation, we know that the VCO output contains two sidebands at $M\omega_{in} \pm \omega_{in}$ [Fig. 9.19(a)]. How does the divider respond to such a spectrum? Since a frequency divider simply divides the input frequency or phase, we can write V_F as

$$V_F(t) = V_0 \cos\left[\frac{1}{M}(M\omega_{in}t + K_{VCO}\int V_m \sin\omega_{in}t dt)\right] \tag{9.10}$$

$$\approx V_0 \cos\omega_{in}t - \frac{K_{VCO}V_m}{2M\omega_{in}} V_0 \cos(\omega_{in} + \omega_{in})t + \frac{K_{VCO}V_m}{2M\omega_{in}} V_0 \cos(\omega_{in} - \omega_{in})t. \tag{9.11}$$

That is, the sidebands maintain their spacing with respect to the carrier after frequency division, but their relative magnitude *falls* by a factor of M. The result is shown in Fig. 9.19(b).

The PLL of Fig. 9.18(b) can also *synthesize* frequencies: if the divider modulus changes by 1, the output frequency changes by ω_{in}. This point forms the basis for the frequency synthesizers studied in Chapter 10.

How does the presence of a feedback divider affect the loop dynamics? In analogy with the op amp circuit of Fig. 9.18(a), we surmise that the weaker feedback leads to a slower response and a larger phase error. We study the response in Problem 9.7 and the phase error in the following example.

Example 9.13

Repeat the analysis of Fig. 9.11 for the PLL of Fig. 9.18(b) and calculate the static phase error.

Solution:

If ω_{in} changes by $\Delta\omega$, ω_{out} must change by $M\Delta\omega$. Such a change translates to a control voltage change equal to $M\Delta\omega/K_{VCO}$ and hence a phase error change of $M\Delta\omega/(K_{VCO}K_{PD})$. As expected, the error is larger by a factor of M.

9.2.6 Drawbacks of Simple PLL

Modern RF synthesizers rarely employ the simple PLL studied here. This is for two reasons. First, Eq. (9.8) imposes a tight relation between the loop stability (ζ) and the corner frequency of the low-pass filter. Recall from Example 9.12 that the ripple on the control line modulates the VCO frequency and must be suppressed by choosing a *low* value for ω_{LPF}. But, a small ω_{LPF} leads to a less stable loop. We seek a PLL topology that does not exhibit this trade-off.

Second, the simple PLL suffers from a limited "acquisition range," e.g., if the VCO frequency and the input frequency are very different at the startup, the loop may never "acquire" lock.[2] Without delving into the process of lock acquisition, we wish to avoid this issue completely so that the PLL always locks.

While not directly relevant to RF synthesizers, the finite static phase error and its variation with the input frequency [Eq. (9.5)] also prove undesirable in some applications. This error can be driven to zero by means of an infinite loop gain—as explained in the next section.

9.3 TYPE-II PLLS

We continue our development by first addressing the second issue mentioned above, namely, the problem of limited acquisition range. While beyond the scope of this book, this limitation arises because *phase* detectors produce little information if they sense *unequal frequencies* at their inputs. We therefore postulate that the acquisition range can be widened

2. Traditional PLLs are characterized by "acquisition range," "pull-in range," "capture range," "lock range," "tracking range," etc. We will soon see that modern PLLs need not deal with these distinctions.

if a *frequency* detector is added to the loop. Of course, we note from Example 9.5 that an FD by itself does not suffice and the loop must eventually lock the phases. Thus, it is desirable to seek a circuit that operates as an FD if its input frequencies are not equal and as a PD if they are. Such a circuit is called a "phase/frequency detector" (PFD).

9.3.1 Phase/Frequency Detectors

Figure 9.20 conceptually shows the operation of a PFD. The circuit produces *two* outputs, Q_A and Q_B, and operates based on the following principles: (1) a rising edge on A yields a rising edge on Q_A (if Q_A is low), and (2) a rising edge on B resets Q_A (if Q_A is high). The circuit is symmetric with respect to A and B (and Q_A and Q_B). We observe from Fig. 9.20(a) that, if $\omega_A > \omega_B$, then Q_A produces pulses while Q_B remains at zero. Conversely, if $\omega_B > \omega_A$, then positive pulses appear at Q_B and $Q_A = 0$. On the other hand, as depicted in Fig. 9.20(b), if $\omega_A = \omega_B$, the circuit generates pulses at either Q_A or Q_B with a width equal to the phase difference between A and B. Thus, the average value of $Q_A - Q_B$ represents the frequency or phase difference.

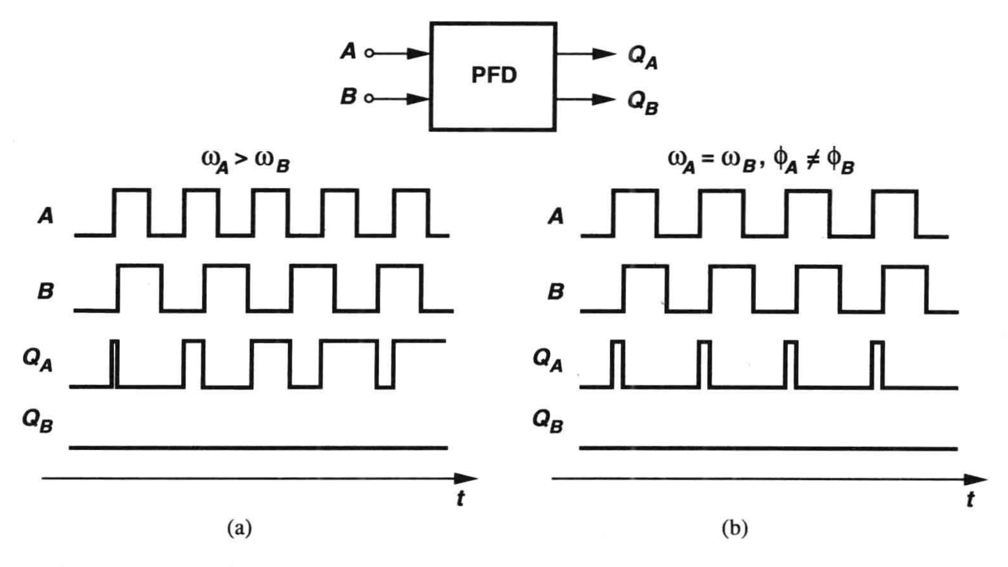


Figure 9.20 *Response of a PFD to inputs with unequal (a) frequencies, or (b) phases.*

To arrive at a circuit implementation of the above idea, we surmise that at least three logical states are necessary: $Q_A = Q_B = 0$; $Q_A = 0$, $Q_B = 1$; and $Q_A = 1$, $Q_B = 0$. Also, to avoid dependence of the output upon the duty cycle of the inputs, the circuit should be realized as an edge-triggered sequential machine. Figure 9.21 shows a state diagram

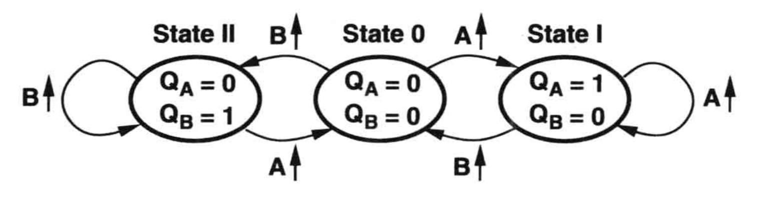


Figure 9.21 *State diagram showing desired operation of PFD.*

summarizing the operation. If the PFD is in state 0, then a transition on A takes it to state I, where $Q_A = 1$, $Q_B = 0$. The circuit remains in this state until a transition occurs on B, upon which the PFD returns to state 0. The switching sequence between states 0 and II is similar.

Figure 9.22 illustrates a logical implementation of the above state machine. The circuit consists of two edge-triggered, resettable D flipflops with their D inputs tied to logical ONE. Signals A and B act as clock inputs of DFF_A and DFF_B, respectively, and the AND gate resets the flipflops if $Q_A = Q_B = 1$. We note that a transition on A forces Q_A to be equal to D input, i.e., a logical ONE. Subsequent transitions on A have no effect. When B goes high, so does Q_B, activating the reset of the flipflops. Thus, Q_A and Q_B are simultaneously high for a duration given by the total delay through the AND gate and the reset path of the flipflops. The reader can show that, if A and B are exactly in-phase, both Q_A and Q_B exhibit these narrow "reset pulses."

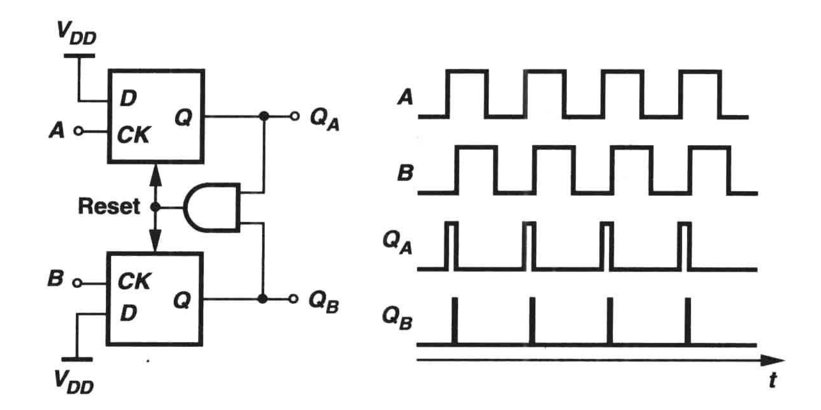


Figure 9.22 *PFD implementation.*

What is the effect of reset pulses on Q_B in Fig. 9.22? Since only the average value of $Q_A - Q_B$ is of interest, these pulses do not interfere with the operation. However, as explained in Section 9.4, the reset pulses introduce a number of errors that tend to increase the ripple on the control voltage.

Each resettable D-flipflop in Fig. 9.22 can be implemented as shown in Fig. 9.23. (Note that no D input is available.) This circuit suffers from a limited speed—a minor issue because in frequency-multiplying PLLs, ω_{in} is typically much lower than ω_{out}. For

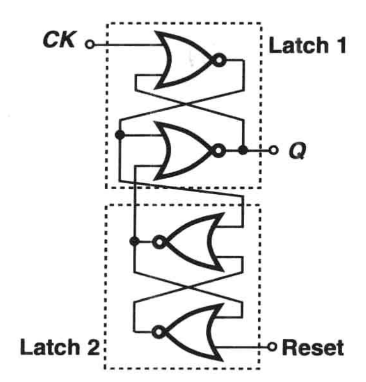


Figure 9.23 *Logical implementation of resettable D flipflop.*

example, in a GSM system, one may choose $\omega_{in} = 2\pi \times (200\,\text{kHz})$ and $\omega_{out} = 2\pi \times (900\,\text{MHz})$. By analyzing the propagation of the reset command in Figs. 9.22 and 9.23, the reader can show that the width of the narrow reset pulses on Q_A and Q_B is equal to three gate delays plus the delay of the AND gate. If the AND gate consists of a NAND gate and an inverter, the pulse width reaches five gate delays. We hereafter assume the reset pulses are five gate delays wide for a zero input phase difference.

The use of a PFD in a phase-locked loop resolves the issue of the limited acquisition range. Shown in Fig. 9.24 is a conceptual realization employing a PFD. The dc content of $Q_A - Q_B$ is extracted by the low-pass filters and amplifier A_1. At the beginning of a transient, the PFD acts as a frequency detector, pushing the VCO frequency toward the input frequency. After the two are sufficiently close, the PFD operates as a phase detector, bringing the loop into phase lock. Note that the polarity of feedback is important here [but not in the simple PLL (Example 9.11)].

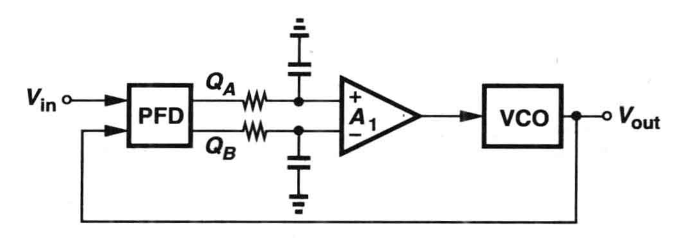


Figure 9.24 *Use of PFD in a type-I PLL.*

We must next address the trade-off between the damping factor and the corner frequency of the loop filter [Eq. (9.8)]. This is accomplished by introducing a "charge pump" (CP) in the loop.

9.3.2 Charge Pumps

A charge pump sinks or sources current for a limited period of time. Depicted in Fig. 9.25 is an example, where switches S_1 and S_2 are controlled by the inputs "Up" and "Down," respectively. A pulse of width ΔT on Up turns S_1 on for ΔT seconds, allowing I_1 to charge C_1. Consequently, V_{out} goes *up* by an amount equal to $\Delta T \cdot I_1/C_1$. Similarly, a pulse

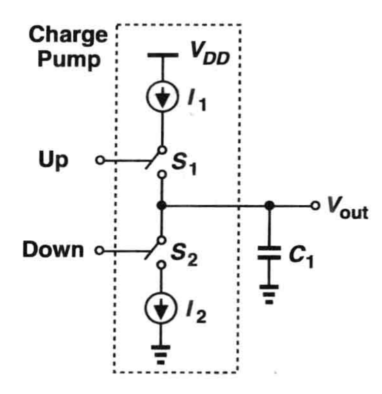


Figure 9.25 *Charge pump.*

on Down yields a drop in V_{out}. Nominally, $I_1 = I_2 = I_p$. Thus, if Up and Down are asserted simultaneously, I_1 simply flows through S_1 and S_2 to I_2, creating no change in V_{out}.

Let us precede the circuit of Fig. 9.25 with a PFD (Fig. 9.26). We note that if, for example, A leads B, then Q_A produces pulses and V_{out} continues to rise. A key point here is that an *arbitrarily small* (constant) phase difference between A and B still turns one switch on—albeit briefly—thereby charging or discharging C_1 and driving V_{out} toward $+\infty$ or $-\infty$—albeit slowly. In other words, the circuit of Fig. 9.26 exhibits an infinite gain, where the gain is defined as the final value of V_{out} divided by the input phase difference. From another perspective, the PFD/CP/C_1 cascade produces a ramp-like output in response to a constant phase difference, displaying the behavior of an integrator.

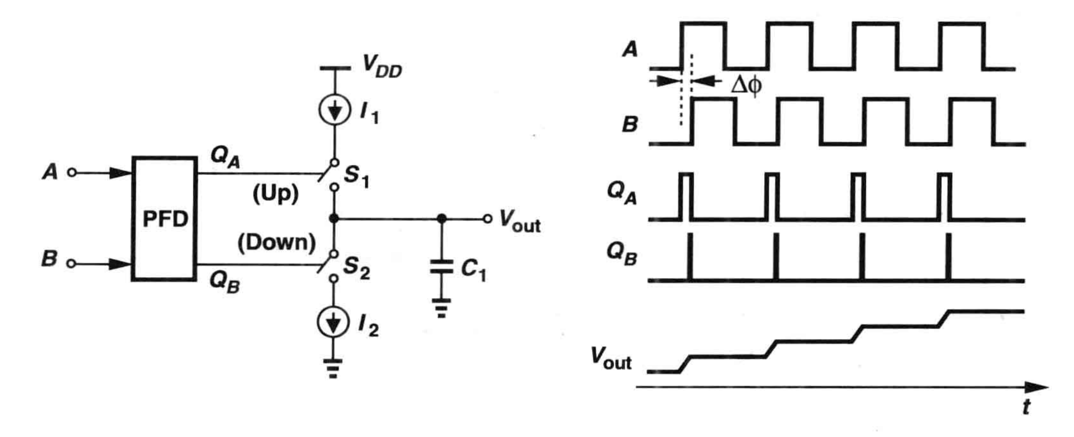


Figure 9.26 *Operation of PFD/CP cascade.*

Example 9.14

We can approximate the PFD/CP circuit of Fig. 9.26 as a current source of some average value driving C_1. Calculate the average value of the current source and the output slope for an input period of T_{in}.

Solution:

For an input phase difference of $\Delta\phi$ rad $= [\Delta\phi/(2\pi)] \times T_{in}$ seconds, the average current is equal to $I_p\Delta\phi/(2\pi)$ and the average slope, $I_p\Delta\phi/(2\pi)/C_1$.

9.3.3 Charge-Pump PLLs

We now construct a PLL using the circuit of Fig. 9.26. Illustrated in Fig. 9.27, such a loop ideally forces the input phase error to zero because, as mentioned in the previous section, a finite error would lead to an *unbounded* value for V_{cont}. To quantify the behavior of this arrangement, we wish to derive the transfer function from ϕ_{in} to ϕ_{out}. Let us first study the transfer function of the PFD/CP/C_1 cascade.

How do we compute this transfer function? We can apply a (phase) step at the input, derive the time-domain output, differentiate it, and compute its Laplace transform [2].

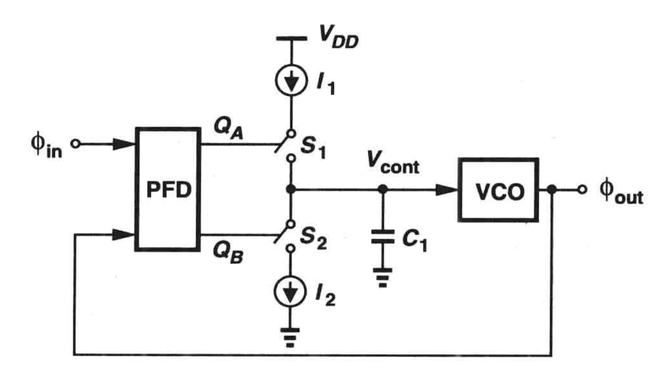


Figure 9.27 *First attempt at constructing a charge-pump PLL.*

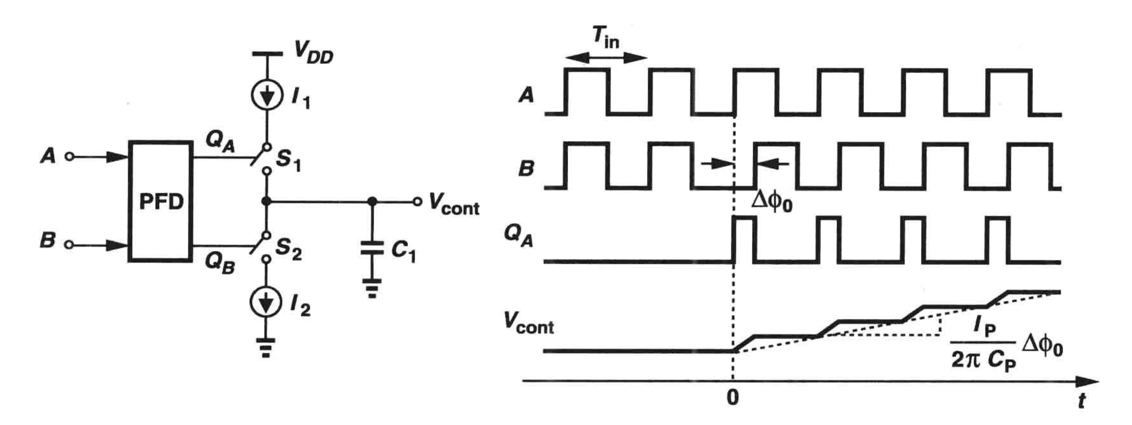


Figure 9.28 *Derivation of the phase step response of PFD/CP/capacitor cascade.*

A phase step simply means a displacement of the zero crossings. As shown in Fig. 9.28, a phase step of $\Delta\phi_0$ at one of the inputs repetitively turns S_1 or S_2 on, monotonically changing the output voltage.

This behavior is similar to that of an integrator. Unfortunately, however, this system is nonlinear: if $\Delta\phi_0$ is doubled, not every point on the "charge-and-hold" output waveform, V_{out}, is doubled (why?). Fortunately, we can approximate this waveform by a ramp—as if the charge pump *continuously* injected current into C_1 (Example 9.14). We call this a "continuous-time (CT) approximation." The change in V_{cont} in every period is equal to

$$\Delta V_{cont} = \frac{\Delta\phi_0}{2\pi} T_{in} \frac{I_p}{C_1}, \tag{9.12}$$

where $[\Delta\phi_0/(2\pi)]T_{in}$ denotes the phase difference in seconds and $I_p = I_1 = I_2$. The slope of the ramp is given by $\Delta V_{cont}/T_{in}$ and hence

$$V_{cont}(t) \approx \frac{\Delta\phi_0}{2\pi} \frac{I_p}{C_1} t u(t). \tag{9.13}$$

Differentiating Eq. (9.13) with respect to time, normalizing to $\Delta\phi_0$, and taking the Laplace transform, we have

$$\frac{V_{cont}}{\Delta\phi}(s) = \frac{I_p}{2\pi C_1}\frac{1}{s}. \tag{9.14}$$

As predicted earlier, the PFD/CP/C_1 cascade operates as an integrator.

Example 9.15

Plot the derivatives of V_{cont} and its ramp approximation in Fig. 9.28 and explain under what condition the derivatives resemble each other.

Solution:

Shown in Fig. 9.29 are the derivatives. The approximation of repetitive pulses by a single step appears less convincing than the approximation of the charge-and-hold waveform by a ramp. Indeed, if a function $f(x)$ can approximate another function $g(x)$, the derivative of $f(x)$ does not necessarily provide a good approximation of the derivative of $g(x)$. Nonetheless, if the time scale of interest is much longer than the input period, we can view the step as an average of the repetitive pulses. Thus, the height of the step is equal to $(I_p/C_1)(\Delta\phi_0/2\pi)$.

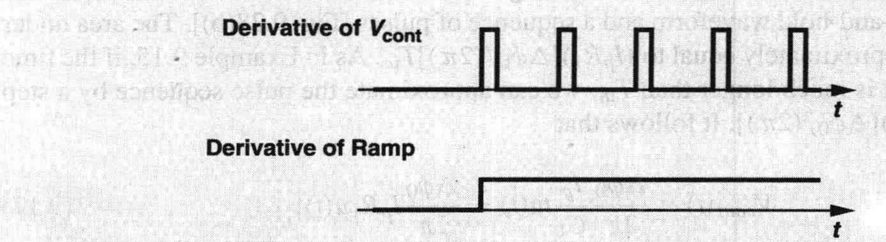


Figure 9.29 *Comparison of the derivatives of the actual phase step response of PFD/CP/capacitor cascade and a ramp.*

From Eq. (9.14), the closed-loop transfer function of the PLL shown in Fig. 9.27 can be expressed as

$$H(s) = \frac{\dfrac{I_p}{2\pi C_1 s}\cdot\dfrac{K_{VCO}}{s}}{1 + \dfrac{I_p}{2\pi C_1 s}\cdot\dfrac{K_{VCO}}{s}} \tag{9.15}$$

$$= \frac{I_p K_{VCO}}{2\pi C_1 s^2 + I_p K_{VCO}}. \tag{9.16}$$

This arrangement is called a type-II PLL because its open-loop transfer function contains two poles at the origin (i.e., two ideal integrators).

Equation (9.16) reveals two poles on the $j\omega$ axis, indicating an oscillatory system. From Example 8.5 for two ideal (lossless) integrators in a loop, we note that the instability is to

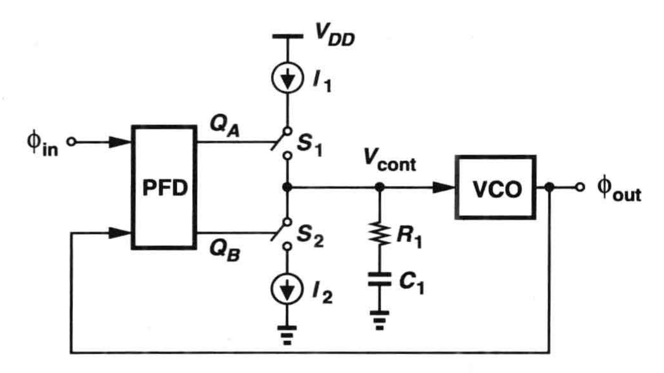


Figure 9.30 *Charge-pump PLL.*

be expected. We thus postulate that if one of the integrators becomes *lossy*, the system can be stabilized. This can be accomplished by inserting a resistor in series with C_1 (Fig. 9.30). The resulting circuit is called a "charge-pump PLL" (CPPLL).

We repeat the analysis illustrated in Fig. 9.28(a) to obtain the new transfer function. As shown in Fig. 9.31, when S_1 or S_2 turns on, V_{cont} jumps by an amount equal to I_pR_1 and subsequently rises or falls linearly with time. When the switch turns off, V_{cont} jumps in the opposite direction, resting at a voltage that is $(I_p/C_1)[\Delta\phi_0/(2\pi)]T_{in}$ volts higher than its value before the switch turned on. The resulting waveform can be viewed as the sum of the original charge-and-hold waveform and a sequence of pulses [Fig. 9.28(b)]. The area under each pulse is approximately equal to $(I_pR_1)[\Delta\phi_0/(2\pi)]T_{in}$. As in Example 9.15, if the time scale of interest is much longer than T_{in}, we can approximate the pulse sequence by a step of height $(I_pR_1)[\Delta\phi_0/(2\pi)]$. It follows that

$$V_{cont}(t) = \frac{\Delta\phi_0}{2\pi}\frac{I_p}{C_1}tu(t) + \frac{\Delta\phi_0}{2\pi}I_pR_1u(t). \tag{9.17}$$

The transfer function of the PFD/CP/filter cascade is therefore given by

$$\frac{V_{cont}}{\Delta\phi}(s) = \frac{I_p}{2\pi}\left(\frac{1}{C_1 s} + R_1\right). \tag{9.18}$$

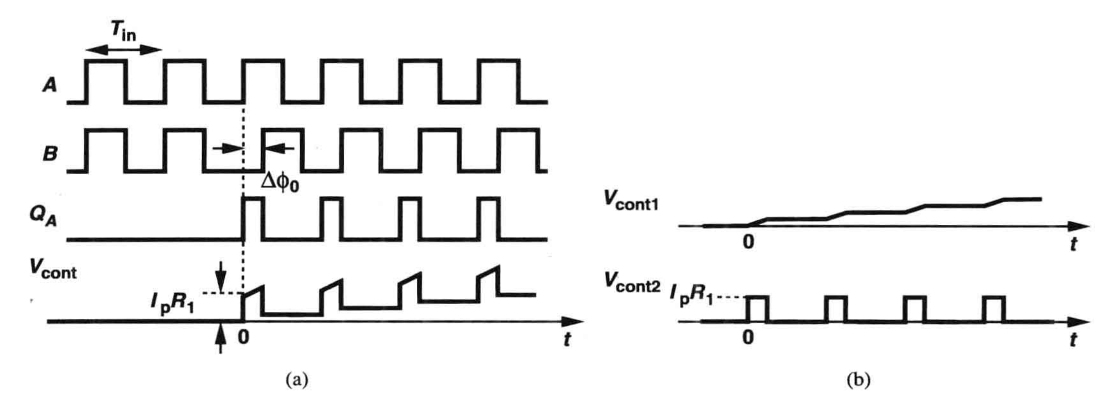


Figure 9.31 *(a) Phase step response of PFD/CP/LPF, (b) decomposition of output waveform into two.*

Equation (9.18) allows us to express the closed-loop transfer function of the PLL shown in Fig. 9.30 as

$$H(s) = \frac{\frac{I_p K_{VCO}}{2\pi C_1}(R_1 C_1 s + 1)}{s^2 + \frac{I_p}{2\pi} K_{VCO} R_1 s + \frac{I_p}{2\pi C_1} K_{VCO}}. \tag{9.19}$$

As with the type-I PLL in Section 9.2, we write the denominator as $s^2 + 2\zeta\omega_n s + \omega_n^2$ and obtain

$$\zeta = \frac{R_1}{2}\sqrt{\frac{I_p C_1 K_{VCO}}{2\pi}} \tag{9.20}$$

$$\omega_n = \sqrt{\frac{I_p K_{VCO}}{2\pi C_1}}. \tag{9.21}$$

Interestingly, as C_1 increases (so as to reduce the ripple on the control voltage), so does ζ—a trend opposite of that observed in type-I PLLs. We have thus removed the trade-off between stability and ripple amplitude. The closed-loop poles are given by

$$\omega_{p1,2} = [-\zeta \pm \sqrt{\zeta^2 - 1}]\omega_n. \tag{9.22}$$

Equation (9.19) also reveals a closed-loop zero at $-\omega_n/(2\zeta)$.

Example 9.16

The PFD implementation of Fig. 9.22 produces two narrow pulses on Q_A and Q_B if A and B have a zero phase difference. Thus, S_1 and S_2 turn on simultaneously for a brief period of time, allowing I_1 to flow to I_2. Does this mean that the CPPLL of Fig. 9.30 is free from ripple?

Solution:

No, in practice, mismatches between I_1 and I_2 and the widths of Up and Down pulses and other imperfections give rise to a finite ripple. We study these effects in Section 9.4.

The transfer function expressed by Eq. (9.18) offers another perspective on stabilization (frequency compensation) of a two-integrator loop. Writing (9.18) as

$$\frac{V_{cont}}{\Delta\phi}(s) = \frac{I_p}{2\pi}\left(\frac{R_1 C_1 s + 1}{C_1 s}\right), \tag{9.23}$$

we can say that a real left-half-plane zero, $\omega_z = -1/(R_1 C_1)$, has been added to the open-loop transfer function, thereby stabilizing the PLL. This point can be better understood by examining the Bode plots of the loop before and after compensation. As shown in Fig. 9.32,

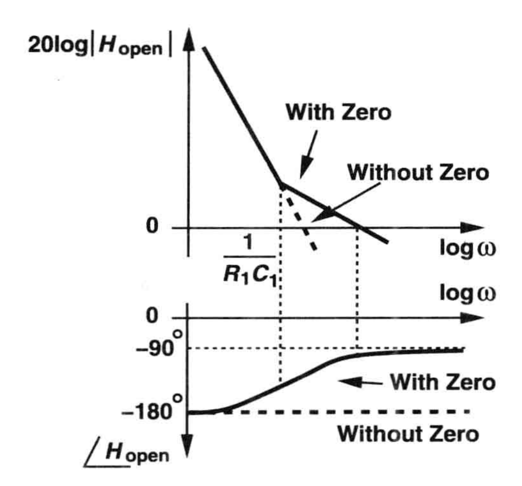


Figure 9.32 *Bode plots of open-loop charge-pump PLL with and without a zero.*

with two ideal integrators, the system has no phase margin, whereas with the zero, both the magnitude and the phase profiles are bent upward, increasing the phase margin.

The behavior illustrated in Fig. 9.32 also explains the dependence of ζ upon K_{VCO} [Eq. (9.20)]. As K_{VCO} decreases, the magnitude plot is shifted down while the phase plot remains unchanged. Thus, the unity-gain frequency moves *closer* to the -180° region, degrading the phase margin. This stands in contrast to the type-I PLL's behavior in Example 9.10.

Suppose during the lock transient, the phase difference is not zero at some point in time. Then, a current of I_p flows through R_1, producing a voltage drop of I_pR_1. In Problem 9.10, we estimate this drop to be $1.6\pi V_{DD}$. (Of course, the CP cannot provide such a large swing.) The key point here is that the control voltage can experience a large jump. We return to this point in Section 9.3.7 and observe that this jump appears even in the locked state, creating significant ripple.

9.3.4 Transient Response

The closed-loop transfer function of the PLL, as expressed by Eq. (9.19), can be used to predict the transient response.

Example 9.17

Plot the magnitude of (9.19) as a function of ω if $\zeta = 1$.

Solution:

The closed loop contains two real coincident poles at $-\omega_n$ and a zero at $-\omega_n/2$. Depicted in Fig. 9.33 $|H|$ begins to rise from unity at $\omega = \omega_n/2$, reaches a peak at $\omega = \omega_n$, returns to unity at $\omega = \sqrt{2}\omega_n$, and continues to fall at a slope of -20 dB/dec thereafter.

Example 9.17 *(Continued)*

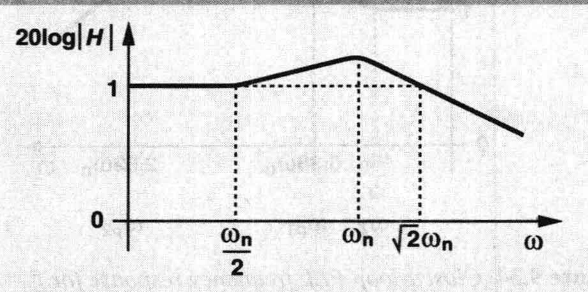


Figure 9.33 *Closed-loop PLL frequency response for two coincident closed-loop poles.*

The inverse Laplace transform of Eq. (9.19) yields the output frequency, $\Delta\omega_{out}$, as a function of time for a frequency step at the input, $\Delta\omega_{in}$:

$$\Delta\omega_{out}(t) = \Delta\omega_{in}u(t) - \Delta\omega_{in}\left[\cos\left(\sqrt{1-\zeta^2}\omega_n t\right) - \frac{\zeta}{\sqrt{1-\zeta^2}}\sin\left(\sqrt{1-\zeta^2}\omega_n t\right)\right]e^{-\zeta\omega_n t}u(t) \quad \zeta < 1 \tag{9.24}$$

$$= \Delta\omega_{in}u(t) - \Delta\omega_{in}(1-\omega_n t)e^{-\zeta\omega_n t}u(t) \quad \zeta = 1 \tag{9.25}$$

$$= \Delta\omega_{in}u(t) - \Delta\omega_{in}\left[\cosh\left(\sqrt{\zeta^2-1}\omega_n t\right) - \frac{\zeta}{\sqrt{\zeta^2-1}}\sinh\left(\sqrt{\zeta^2-1}\omega_n t\right)\right]e^{-\zeta\omega_n t}u(t) \quad \zeta > 1. \tag{9.26}$$

Since the response decays exponentially, we may call $1/(\zeta\omega_n)$ the "time constant" of the loop, but, as explained below, that is not an accurate statement. Note that (9.24) can be simplified if we assume $\zeta/\sqrt{1-\zeta^2} = \tan\psi$ (i.e., $\zeta = \sin\psi$):

$$\Delta\omega_{out}(t) = \Delta\omega_{in}u(t) - \frac{\Delta\omega_{in}}{\sqrt{1-\zeta^2}}\cos\left(\sqrt{1-\zeta^2}\omega_n t + \psi\right)e^{-\zeta\omega_n t}u(t) \quad \zeta < 1. \tag{9.27}$$

From (9.20) and (9.21), the time constant of the loop is expressed as

$$\frac{1}{\zeta\omega_n} = \frac{4\pi}{R_1 I_p K_{VCO}}. \tag{9.28}$$

This quantity (or its inverse) serves as a measure of the settling speed of the loop if ζ is in the vicinity of unity.

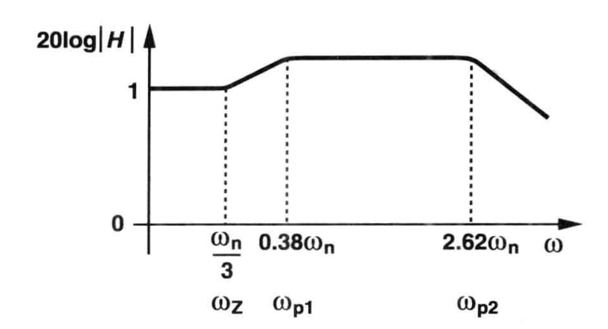


Figure 9.34 *Closed-loop PLL frequency response for* $\zeta = 1.5$.

What happens as ζ well exceeds unity? If $\zeta^2 \gg 1$, then $\sqrt{\zeta^2 - 1} \approx \zeta[1 - 1/(2\zeta^2)] = \zeta - 1/(2\zeta)$, and Eq. (9.22) reduces to

$$\omega_{p1} \approx -\frac{1}{2\zeta}\omega_n = -\frac{1}{R_1 C_1} \tag{9.29}$$

$$\omega_{p2} \approx -2\zeta\omega_n = -\frac{R_1 I_p K_{VCO}}{2\pi}. \tag{9.30}$$

Note that $\omega_{p1}/\omega_{p2} \approx 4\zeta^2 \gg 1$. Does this mean ω_{p2} becomes a dominant pole? No, interestingly, the *zero* is also located at $-\omega_n/(2\zeta)$, cancelling the effect of ω_{p2}. Thus, for large ζ^2, the loop approaches a one-pole system having a time constant of $1/|\omega_{p1}| = 1/(2\zeta\omega_n)$. Figure 9.34 plots $|H|$ for $\zeta = 1.5$, indicating that even this value of ζ leads to an approximately one-pole response because ω_z and ω_{p1} are relatively close.

Example 9.18

A student has encountered an inconsistency in our derivations. We concluded above that the loop time constant is approximately equal to $1/(2\zeta\omega_n)$ for $\zeta^2 \gg 1$, but (9.24)-(9.26) evidently imply a time constant of $1/(\zeta\omega_n)$. Explain the cause of this inconsistency.

Solution:

For $\zeta^2 \gg 1$, we have $\zeta/\sqrt{\zeta^2 - 1} \approx 1$. Since $\cosh x - \sinh x = e^{-x}$, we rewrite Eq. (9.26) as

$$\Delta\omega_{out}(t) = \Delta\omega_{in}u(t) - N\Delta\omega_{in}(e^{-\zeta\omega_n t})e^{-\zeta\omega_n t}u(t) \quad \zeta^2 \gg 1. \tag{9.31}$$

Thus, the time constant of the loop is indeed equal to $1/(2\zeta\omega_n)$. More generally, we say that with typical values of ζ, the loop time constant lies between $1/(\zeta\omega_n)$ and $1/(2\zeta\omega_n)$.

9.3.5 Limitations of Continuous-Time Approximation

Charge-pump PLLs are inherently discrete-time (DT) systems because the charge pump turns off for part of the period and breaks the loop. In the derivation of the CPPLL transfer function, we have made two continuous-time approximations: the charge-and-hold waveform in Fig. 9.28 is represented by a ramp, and the series of pulses in Fig. 9.31(b) is

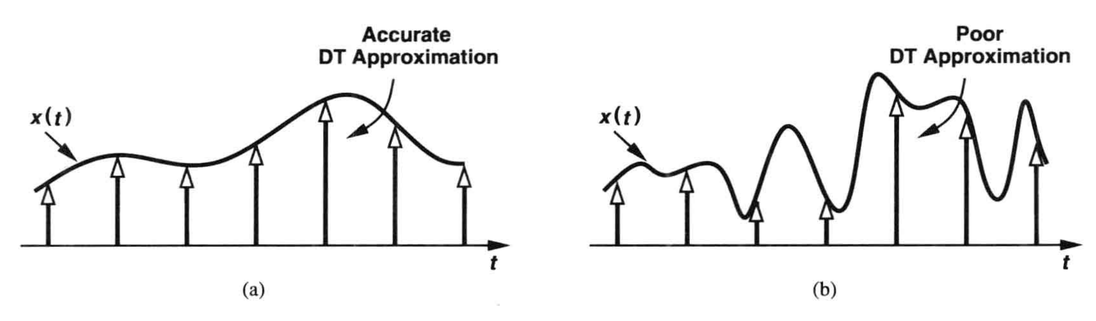


Figure 9.35 *(a) Accurate and (b) poor discrete-time approximations of continuous-time waveforms.*

modeled by a step. These approximations hold only if the time "granularities" inherent in the original waveforms are very small with respect to the time scales of interest. To better understand this point, let us consider the *inverse* problem, namely, the approximation of a CT waveform by a DT counterpart. As illustrated in Fig. 9.35, the approximation holds well if the CT waveform changes little from one clock cycle to the next, but loses its accuracy if the CT waveform experiences fast changes.

These observations reveal that CPPLLs obey the transfer function of Eq. (9.19) only if their internal states (the control voltage and the VCO phase) do not change rapidly from one input cycle to the next. This occurs if the loop time constant is much longer than the input period. Indeed, this point plays an important role in the design procedure of PLLs (Section 9.7). Discrete-time analyses of CPPLLs can be found in [3], but in practice, loops that are not sufficiently slow exhibit an underdamped behavior or may simply not lock. The CT approximation therefore proves adequate in most practical cases.

9.3.6 Frequency-Multiplying CPPLL

As explained in Section 9.2.5 and illustrated in Fig. 9.18(b), a PLL containing a divider of modulus M in its feedback path multiplies the input frequency by a factor of M. We wish to formulate the dynamics of a type-II frequency-multiplying PLL. We simply consider the product of (9.23) and K_{VCO}/s as the forward transfer function and $1/M$ as the feedback factor, arriving at

$$H(s) = \frac{\dfrac{I_p K_{VCO}}{2\pi C_1}(R_1 C_1 s + 1)}{s^2 + \dfrac{I_p}{2\pi}\dfrac{K_{VCO}}{M}R_1 s + \dfrac{I_p}{2\pi C_1}\dfrac{K_{VCO}}{M}}. \tag{9.32}$$

The denominator is similar to that of Eq. (9.19), except that K_{VCO} is divided by M. Thus, (9.20) and (9.21) can be respectively modified to

$$\zeta = \frac{R_1}{2}\sqrt{\frac{I_p C_1}{2\pi}\frac{K_{VCO}}{M}} \tag{9.33}$$

$$\omega_n = \sqrt{\frac{I_p}{2\pi C_1}\frac{K_{VCO}}{M}}. \tag{9.34}$$

As can be seen in Fig. 9.32, the division of K_{VCO} by M makes the loop *less* stable (why?), requiring that I_p and/or C_1 be larger. We can rewrite (9.32) as

$$H(s) = M\frac{2\zeta\omega_n s + \omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}. \tag{9.35}$$

Example 9.19

The input to a multiplying PLL is a sinusoid with two small "close-in" FM sidebands, i.e., the modulation frequency is relatively low. Determine the output spectrum of the PLL.

Solution:

The input can be expressed as

$$V_{in}(t) = V_0 \cos(\omega_{in}t + a\int \cos\omega_m t dt) \tag{9.36}$$

$$= V_0 \cos\left(\omega_{in}t + \frac{a}{\omega_m}\sin\omega_m t\right). \tag{9.37}$$

Since the sidebands are small, the narrowband FM approximation applies and the magnitude of the input sidebands normalized to the carrier amplitude is equal to $a/(2\omega_m)$. Since $\sin\omega_m t$ modulates the phase of the input *slowly*, we let $s \to 0$ in (9.35),

$$\frac{\phi_{out}}{\phi_{in}}(s \approx 0) = M, \tag{9.38}$$

an expected result because frequency multiplication and phase multiplication are synonymous. The output phase modulation is therefore M times the input phase modulation:

$$V_{out}(t) = V_1 \cos\left(M\omega_{in}t + \frac{Ma}{\omega_m}\sin\omega_m t\right). \tag{9.39}$$

In other words, the relative magnitude of the sidebands *grows* by a factor of M, but their spacing with respect to the carrier remains constant (Fig. 9.36). This behavior is the reverse of that observed for frequency division in Example 9.12.

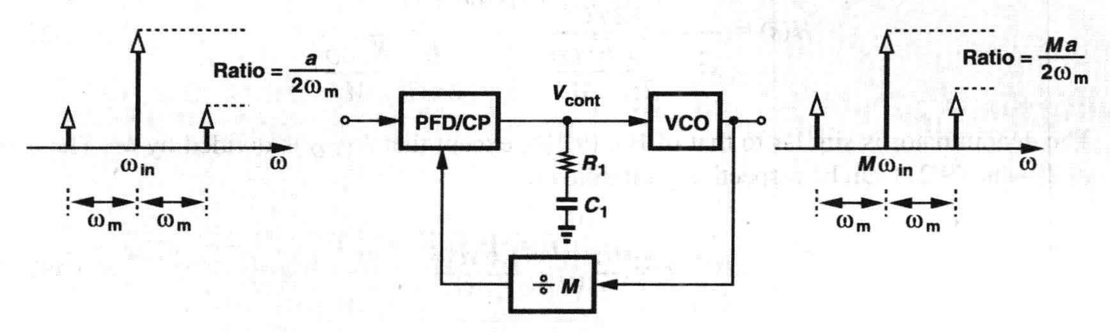


Figure 9.36 *Amplification of sidebands in a frequency-multiplying PLL.*

It is important to recognize that the above example applies to only *slow* phase or frequency modulation at the PLL input such that the output tracks the variation faithfully. For faster modulations, the output phase is an attenuated version of the input and subjected to Eq. (9.32).

9.3.7 Higher-Order Loops

The loop filter consisting of R_1 and C_1 in Fig. 9.30 proves inadequate because, even in the locked condition, it does not suppress the ripple sufficiently. For example, suppose, in the locked condition, the Up and Down pulses arrive every T_{in} seconds with a small skew due to propagation mismatches within the PFD (Fig. 9.37). Consequently, one switch turns on earlier than the other, allowing its corresponding current source to flow through R_1 and generate an instantaneous change of I_pR_1 in the control voltage. On the falling edge of the Up and Down pulses, the reverse happens. The ripple thus consists of positive and negative pulses of amplitude I_pR_1 occurring every T_{in} seconds. Since I_pR_1 is quite large (even higher than the supply voltage!),[3] additional means of reducing the ripple become necessary.

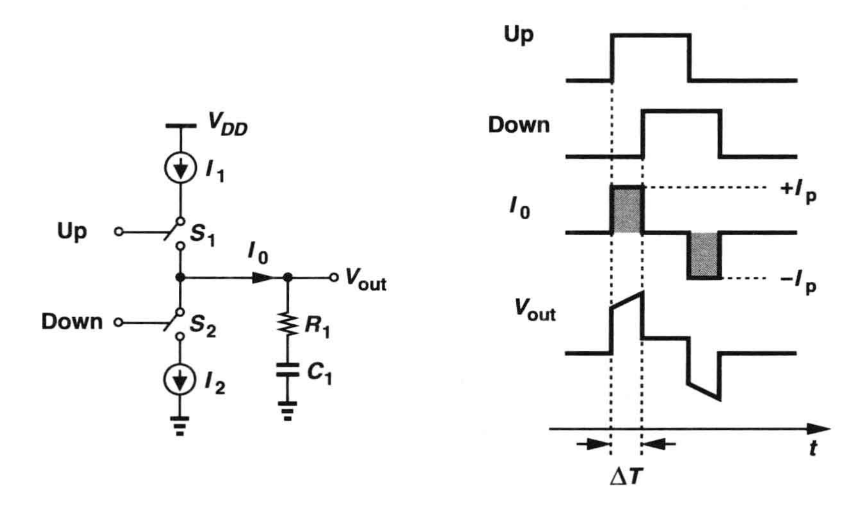


Figure 9.37 *Effect of skew between Up and Down pulses.*

A common approach to lowering the ripple is to tie a capacitor directly from the control line to ground. Illustrated in Fig. 9.38, the idea is to provide a low-impedance path for the unwanted charge pump output. That is, a current pulse of width ΔT produced by the CP initially flows through C_2, leading to a change of $(I_p/C_2)\Delta T$ in V_{cont}. (Since typically $R_1C_2 \gg \Delta T$, the voltage change can be approximated by a ramp.) After the CP turns off, C_2 begins to share its charge with C_1 through R_1, causing an exponential decay in V_{cont} with a time constant of R_1C_{eq}, where $C_{eq} = C_1C_2/(C_1 + C_2)$. Of course, it is hoped that C_2 can be large enough to yield a small ripple.

How large can C_2 be? The current-to-voltage conversion impedance provided by the loop filter has changed from $R_1 + (C_1s)^{-1}$ to $[R_1 + (C_1s)^{-1}]||(C_2s)^{-1}$, presenting an

3. In Problem 9.10, we estimate I_pR_1.

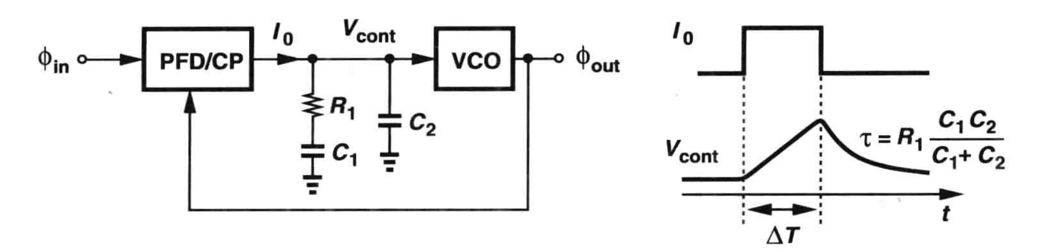


Figure 9.38 *Addition of second capacitor to loop filter.*

additional pole at $(R_1 C_{eq})^{-1}$ and degrading the loop stability. We must therefore compute the phase margin before and after addition of C_2. As shown in Appendix I,

$$\text{PM} \approx \tan^{-1}(4\zeta^2) - \tan^{-1}\left(4\zeta^2 \frac{C_{eq}}{C_1}\right), \tag{9.40}$$

where ζ is chosen equal to $0.5\sqrt[4]{C_1/C_{eq}}$ to maximize PM. For example, if $C_2 = 0.2C_1$, then $(R_1 C_{eq})^{-1} = 6\omega_z$, $\zeta \approx 0.783$, and the phase margin falls from 76° to 46°. Simulations of PLLs indicate that this estimate is somewhat pessimistic and $C_2 \leq 0.2C_1$ is a reasonable choice in most cases. We therefore choose $\zeta = 0.8$-1 and $C_2 \approx 0.2C_1$ in typical designs.[4]

Unfortunately, with C_2 present, R_1 cannot be arbitrarily large. In fact, if R_1 is so large that the series combination of R_1 and C_1 is overwhelmed by C_2, then the PLL reduces to the system shown in Fig. 9.27 and characterized by Eq. (9.16). An upper bound derived for R_1 in Appendix I is as follows:

$$R_1^2 \leq \frac{2\pi}{I_p K_{VCO} C_{eq}}. \tag{9.41}$$

Example 9.20

Consider the two filter/VCO topologies shown in Fig. 9.39 and explain which one is preferable with respect to supply noise.

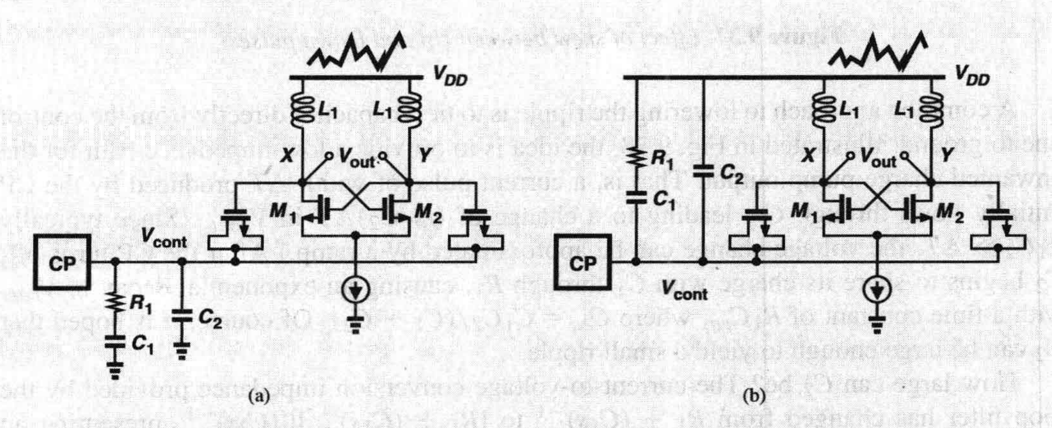


Figure 9.39 *Loop filter referenced to (a) ground, and (b) V_{DD}.*

4. Note that ζ is still the damping factor of the original second-order loop.

Example 9.20 *(Continued)*

Solution:

In Fig. 9.39(a), the loop filter is "referenced" to ground, whereas the voltage across the varactors is referenced to V_{DD}. Since C_1 and C_2 are much greater than the capacitance of the varactors, V_{cont} remains relatively constant and noise on V_{DD} modulates the value of the varactors. In Fig. 9.39(b), on the other hand, the loop filter and the varactors are referenced to the same "plane," namely, V_{DD}. Thus, noise on V_{DD} negligibly modulates the voltage across the varactors. In essence, the loop filter "bootstraps" V_{cont} to V_{DD}, allowing the former to track the latter. This topology is therefore preferable. This principle should be observed for the interface between the loop filter and the VCO in any PLL design.

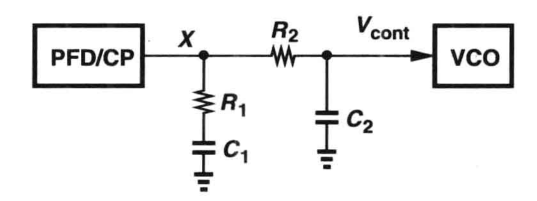


Figure 9.40 *Alternative second-order loop filter.*

Another loop filter that can reduce the ripple is shown in Fig. 9.40. Here, the ripple at node X may be large, but it is suppressed as it travels through the low-pass filter consisting of R_2 and C_2. If $|R_2 + (C_2s)^{-1}| \gg |R_1 + (C_1s)^{-1}|$ at the frequencies of interest, then the additional pole is given by $(R_2C_2)^{-1}$. Following the analysis in Appendix I, the reader can prove that

$$\text{PM} \approx \tan^{-1}(4\zeta^2) - \tan^{-1}\left(4\zeta^2\frac{R_2C_2}{R_1C_1}\right). \tag{9.42}$$

Thus, $(R_2C_2)^{-1}$ must remain 5 to 10 times higher than ω_z so as to yield a reasonable phase margin.

9.4 PFD/CP NONIDEALITIES

Our study of PLLs in the previous sections has provided a detailed understanding of their basic operation but has neglected various imperfections. In this section, we analyze the effect of nonidealities in the PFD/CP cascade. We also present circuit techniques that combat some of these effects.

9.4.1 Up and Down Skew and Width Mismatch

The Up and Down pulses produced by the PFD may arrive at different times. As explained in Section 9.3.7 and illustrated in Fig. 9.37, an arrival time mismatch of ΔT translates to two current pulses of width ΔT, height I_p, and opposite polarities that are injected by the charge pump at each phase comparison instant. Owing to the short time scales associated

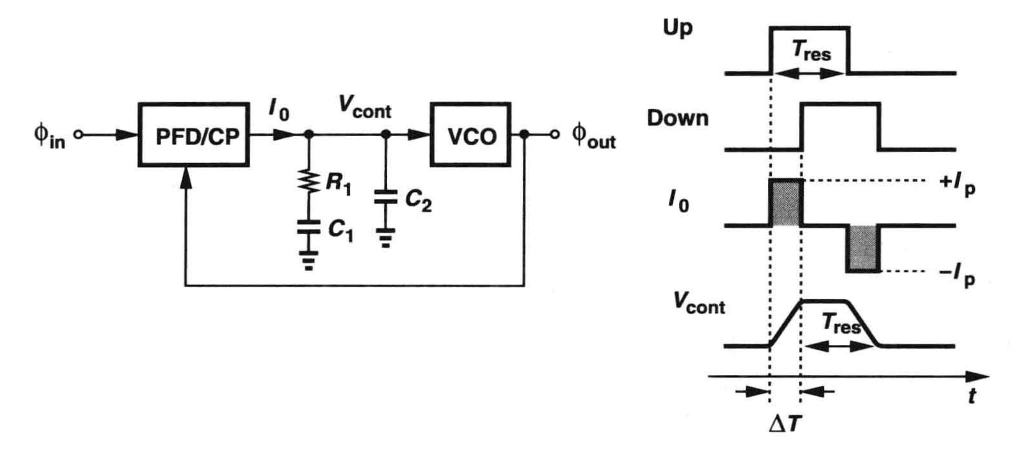


Figure 9.41 *Effect of Up and Down skew on V_{cont} for a second-order filter.*

with these pulses, only C_2 in Fig. 9.38 acts as a storage element, producing a pulse on the control line (Fig. 9.41). The width of the pulse is equal to the width of the reset pulses, T_{res} (about 5 gate delays for the PFD implementation of Figs. 9.22 and 9.23), plus ΔT. The height of the pulse is equal to $\Delta T I_p/C_2$.

Example 9.21

Approximating the pulses on the control line by impulses, determine the magnitude of the resulting sidebands at the output of the VCO.

Solution:

The area under each pulse is approximately given by $(\Delta T I_p/C_2)T_{res}$ if $T_{res} \gg 2\Delta T$. The Fourier transform of the sequence therefore contains impulses at the multiples of the input frequency, $f_{in} = 1/T_{in}$, with an amplitude of $(\Delta T I_p/C_2)T_{res}/T_{in}$ (Fig. 9.42). The two impulses at $\pm 1/T_{in}$ correspond to a sinusoid having a peak amplitude of $2\Delta T I_p C_2 T_{res}/T_{in}$. If the narrowband FM approximation holds, we conclude that the relative magnitude of the sidebands at $f_c \pm f_{in}$ at the VCO output is given by

$$\frac{A_{side}}{A_{carrier}} = \frac{1}{2\pi}\frac{\Delta T I_p}{C_2} T_{res} K_{VCO}. \tag{9.43}$$

Figure 9.42 *Spectrum of ripple on control voltage.*

Sidebands at $f_c \pm n f_{in}$ are scaled down by a factor of n. An interesting and useful observation here is that Eq. (9.43) is independent of f_{in}.

The reader may wonder how the Up and Down pulses may arrive at different times. In addition to random propagation time mismatches with the PFD, the interface between the PFD and the charge pump may also introduce a systematic skew. For example, consider the arrangement shown in Fig. 9.43(a), where the charge pump is implemented by M_1-M_4. Since S_1 is realized by a PMOS device, the corresponding PFD output, Q_A, must be inverted so that M_1 is on when Q_A is high. The delay of the inverter thus creates a skew between the Up and Down pulses. To alleviate this issue, a transmission gate can be inserted in the *Down* pulse path so as to replicate the delay of the inverter [Fig. 9.43(b)].[5]

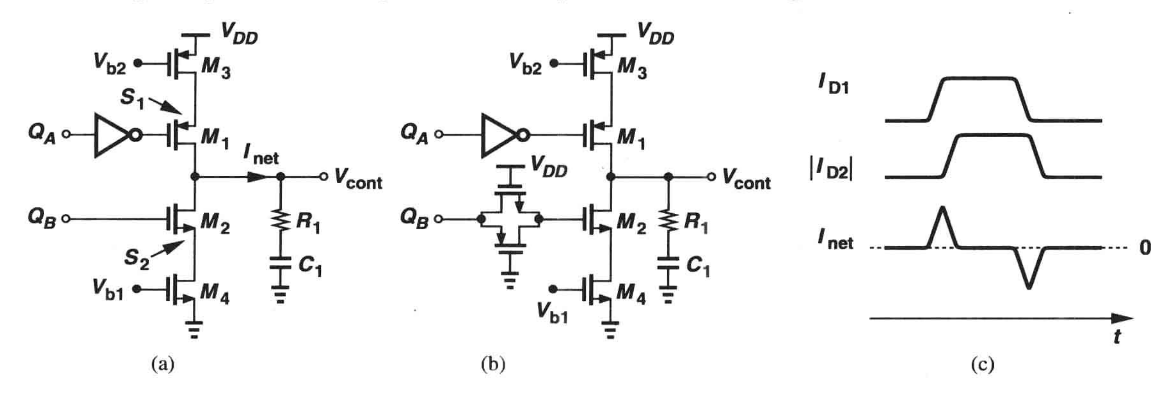


Figure 9.43 *(a) Skew between Up and Down pulses as a result of additional inverter, (b) skew compensation by a pass gate, (c) current waveforms showing effect of skew.*

Does perfect alignment of Up and Down pulses in Fig. 9.43(b) suffice? Not necessarily; the *currents* produced by the PMOS and NMOS sections of the charge pump may still suffer from skews. This is because the time instants at which M_1 and M_2 turn on and off may not be aligned. In other words, the quantity of interest is in fact the skew between the Up and Down *current* waveforms, I_{D1} and I_{D2} in Fig. 9.43(c), respectively, or, ultimately, the net current injected into the loop filter, $I_{net} = I_{D1} - I_{D2}$. In the design of the PFD/CP combination, I_{net} must be minimized in amplitude and in duration.

Example 9.22

What is the effect of mismatch between the *widths* of the Up and Down pulses?

Solution:

Illustrated in Fig. 9.44(a) for the case of Down narrower than Up, this condition may suggest that a pulse of current is injected into the loop filter at each phase comparison instant. However, such periodic injection would continue to increase (or decrease) V_{cont} with no bound. The PLL thus creates a *phase offset* as shown in Fig. 9.44(b) such that the Down pulse becomes as wide as the Up pulse. Consequently, the net current injected into the filter consists of two pulses of equal and opposite areas. For an original width mismatch of ΔT, Eq. (9.43) applies here as well.

(Continues)

5. The skew is not completely cancelled because the capacitance seen by Q_B may be different from that seen by Q_A.

Example 9.22 *(Continued)*

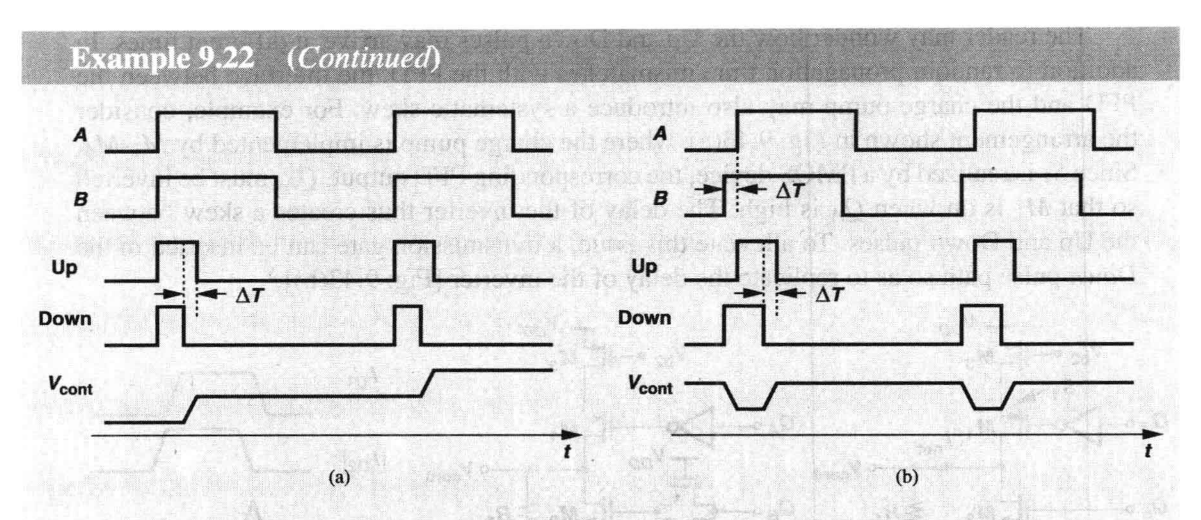


Figure 9.44 *Effect of Up and Down width mismatch: (a) initial response, (b) steady-state response.*

9.4.2 Voltage Compliance

Recall from Chapter 8 that we wish to maximize the tuning range of VCOs while maintaining a moderate value for K_{VCO}. It is therefore desirable to design the charge pump so that it produces minimum and maximum voltages as close to the supply rails as possible. In the simple charge pump of Fig. 9.43(b), each current source requires a minimum drain-source voltage and each switch sustains a voltage drop. We say the output compliance is equal to V_{DD} minus two overdrive voltages and two switch drops. To maximize the output compliance, wide devices must be employed, but at the cost of exacerbating some of the issues described below.

9.4.3 Charge Injection and Clock Feedthrough

We now turn our attention to the imperfections introduced by the charge pump. We assume the simple CP implementation of Fig. 9.43(b) for now. The switching transistors, M_1 and M_2, carry a certain amount of mobile charge in their inversion layers when they are on. This charge is expressed as

$$|Q_{ch}| = WLC_{ox}|V_{GS} - V_{TH}|. \tag{9.44}$$

As the switches turn on, they absorb this charge and as they turn off, they dispel this charge, in both cases through their source and drain terminals. Since M_1 and M_2 generally have different dimensions and overdrive voltages, they do not cancel each other's charge absorption or injection, thereby disturbing the control voltage at both turn-on and turn-off points [Fig. 9.45(a)]. We hereafter refer to this effect as charge injection and consider it when switches turn off, bearing in mind that charge absorption plays a similar role.

Another effect relates to the gate-drain overlap capacitance of the switches. As shown in Fig. 9.45(b), the Up and Down pulses couple through C_{GD1} and C_{GD2}, respectively, and

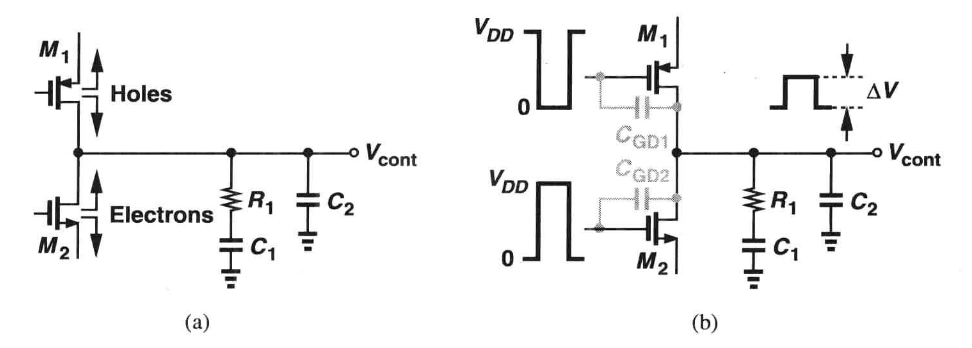


Figure 9.45 *(a) Channel charge injection, and (b) clock feedthrough in a charge pump.*

reach V_{cont}. Since R_1C_1 is quite long, only C_2 attenuates this "clock feedthrough" initially

$$\Delta V = \frac{C_{GD1} - C_{GD2}}{C_{GD1} + C_{GD2} + C_2} V_{DD}. \tag{9.45}$$

After the charge pump turns off, charge sharing between C_2 and C_1 reduces this voltage to

$$\Delta V' = \frac{C_{GD1} - C_{GD2}}{C_{GD1} + C_{GD2} + C_2 + C_1} V_{DD}. \tag{9.46}$$

A number of techniques can reduce the effect of charge injection and clock feedthrough. Depicted in Fig. 9.46(a), one approach places the switches near the supply rails [4] so that the feedthrough is somewhat attenuated by the total capacitance seen from X and Y to ground before disturbing the source voltage of M_3 and M_4. Charge injection, however, persists because M_3 and M_4 must still dispel their charge when they turn off. This approach is called "source switching" because the switches are tied to the sources of M_3 and M_4.

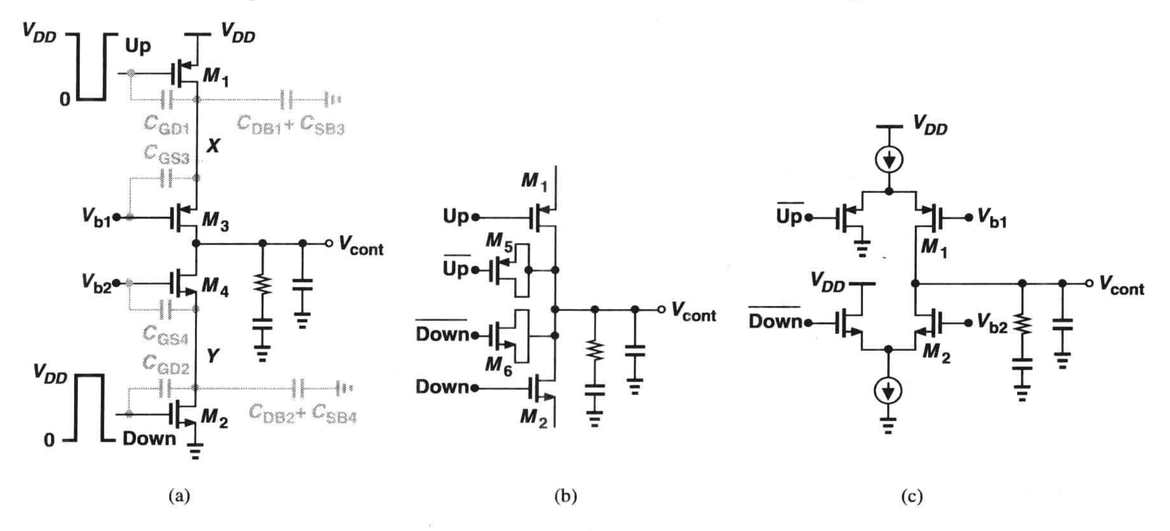


Figure 9.46 *Improved charge pumps: (a) source-switched CP, (b) use of dummy switches, (c) use of differential pairs.*

Another method incorporates "dummy" switches to suppress both effects [5]. Illustrated in Fig. 9.46(b), the idea is to add transistors configured as capacitors and driven by

the complements of Up and Down pulses. The reader can prove that, if $W_5 = 0.5W_1$ and $W_6 = 0.5W_2$, then the clock feedthrough of each switch is cancelled. The charge injection is also cancelled if, additionally, the charge of each switch splits equally between its source and drain terminals. Since this condition is difficult to guarantee, charge injection may be only partially removed.

Figure 9.46(c) shows another arrangement, where the Up and Down currents are created by differential pairs. If V_{b1} and V_{b2} provide a low impedance at the gates of M_1 and M_2, respectively, then $\overline{Up}$ and $\overline{Down}$ find no feedthrough path to the output. However, the charge injection mismatch between M_1 and M_2 remains uncorrected.

9.4.4 Random Mismatch between Up and Down Currents

The two current sources in a charge pump inevitably suffer from random mismatches. Figure 9.47(a) shows an example, where I_{REF} is copied onto M_4 and M_6 and I_{D6} onto M_3. We note that mismatches between M_4 and M_6 and between M_3 and M_5 manifest themselves in the Up and Down currents. How does the PLL react to this mismatch? If, as depicted in Fig. 9.47(b), the Up and Down pulses remain aligned, then a net positive (or negative) current is injected into the loop filter, yielding an unbounded control voltage (in a manner similar to Example 9.22). The loop must therefore develop a phase offset such that the *smaller* current lasts *longer* [Fig. 9.47(c)]. For a mismatch of ΔI, the net current is zero if

$$I_p \cdot \Delta T = \Delta I \cdot T_{res}, \tag{9.47}$$

where I_p denotes the mean current. Thus,

$$\Delta T = T_{res} \frac{\Delta I}{I_p}. \tag{9.48}$$

The ripple amplitude is equal to $\Delta T \cdot I_p / C_2 = T_{res} \Delta I / C_2$.

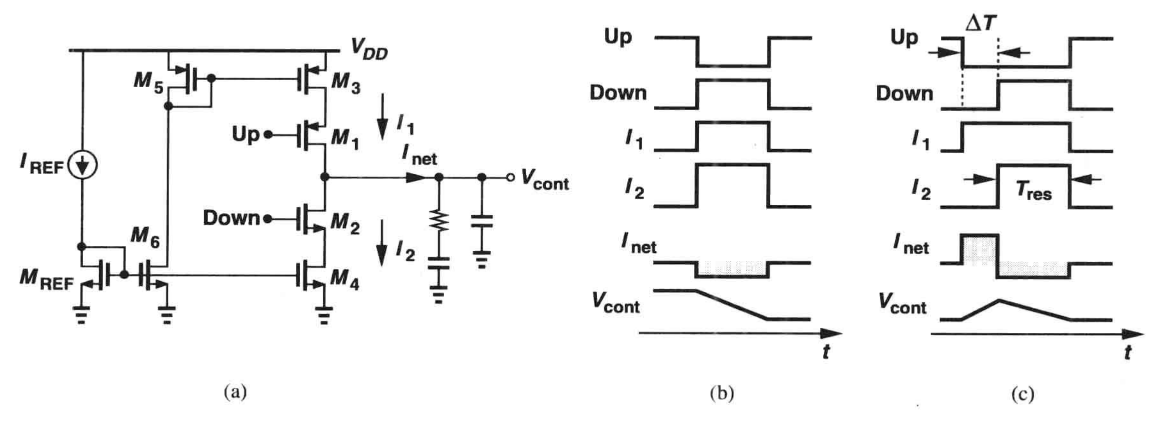


Figure 9.47 *(a) Simple CP realization, (b) initial response to Up and Down current mismatch, (c) steady-state response to Up and Down current mismatch.*

How does the ripple due to Up and Down skew compare with that due to current mismatch? As derived earlier, the former has an amplitude equal to $\Delta T \cdot I_p / C_2$. Thus, one is proportional to the skew times the *entire* charge pump current whereas the other

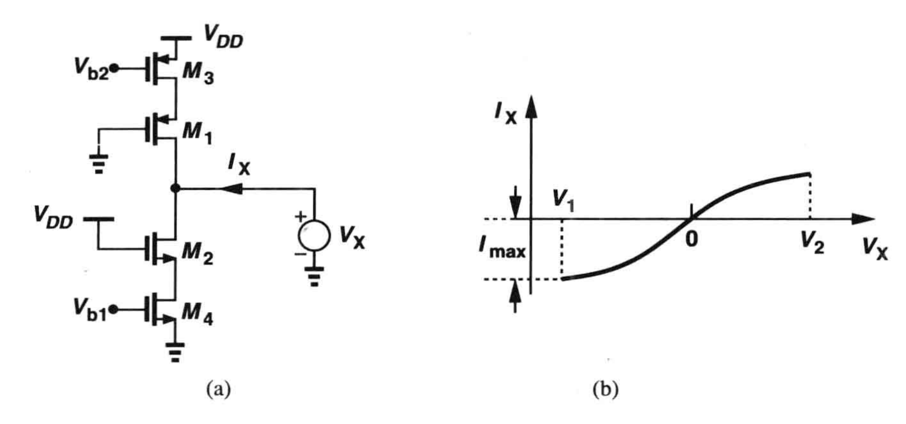


Figure 9.48 *(a) Charge pump configured to measure effect of channel-length modulation, (b) behavior of* I_X.

is proportional to the reset pulsewidth times the current *mismatch*. The two may thus be comparable.

The random mismatch between the Up and Down currents can be reduced by enlarging the current-source transistors. Recall from analog design that as the device area increases, mismatches experience greater spatial averaging. For example, doubling the area of a transistor—equivalent to placing two transistors in parallel—reduces the threshold voltage mismatch by a factor of $\sqrt{2}$. However, larger transistors suffer from a greater amount of charge injection and clock feedthrough.

9.4.5 Channel-Length Modulation

The Up and Down currents also incur mismatch due to channel-length modulation of the current sources; i.e., different output voltages inevitably lead to opposite changes in the drain-source voltages of the current sources, thereby creating a larger mismatch.

In order to quantify the effect of channel-length modulation, we test the charge pump as shown in Fig. 9.48(a). Both switches are on and the output voltage is swept across the compliance range. In the ideal case, $I_X = 0$ for the entire range, but in reality, I_X varies as shown in Fig. 9.48(b) because the PMOS or NMOS source carries a larger current than the other. The maximum departure of I_X from zero, I_{max}, divided by the nominal value of I_p quantifies the effect of channel-length modulation. With short-channel devices, this ratio may reach 30-48%.

Example 9.23

The phase offset of a CPPLL varies with the output frequency. Explain why.

Solution:

At each output frequency and hence at each control voltage, channel-length modulation introduces a certain mismatch between the Up and Down currents [Fig. 9.48(b)]. As implied

(Continues)

Example 9.23 *(Continued)*

by Eq. (9.48), this mismatch is normalized to I_p and multiplied by T_{res} to yield the phase offset. The general behavior is sketched in Fig. 9.49.

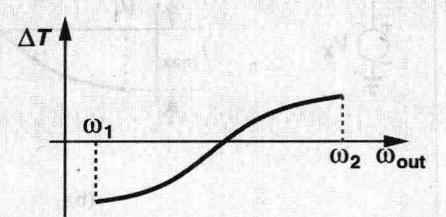


Figure 9.49 *Variation of phase offset with frequency.*

While the phase offset or its variation is not critical in RF synthesis, the resulting ripple is. That is, channel-length modulation must be small enough to produce a tolerable ripple amplitude ($= T_{res}\Delta I/C_2$). Longer transistors can alleviate this effect, but in practice it may be difficult to achieve sufficiently small ΔI. For this reason, a number of circuit techniques have been devised to deal with channel-length modulation.

9.4.6 Circuit Techniques

It is possible to raise the output impedance of the current sources through the use of "regulated cascodes" [6]. Figure 9.50(a) illustrates such a structure, where an "auxiliary amplifier," A_0, senses V_P and adjusts the gate voltage of M_1 so as to maintain V_P close to V_b and hence the current through R_S, I_X, relatively constant. As a result, the output

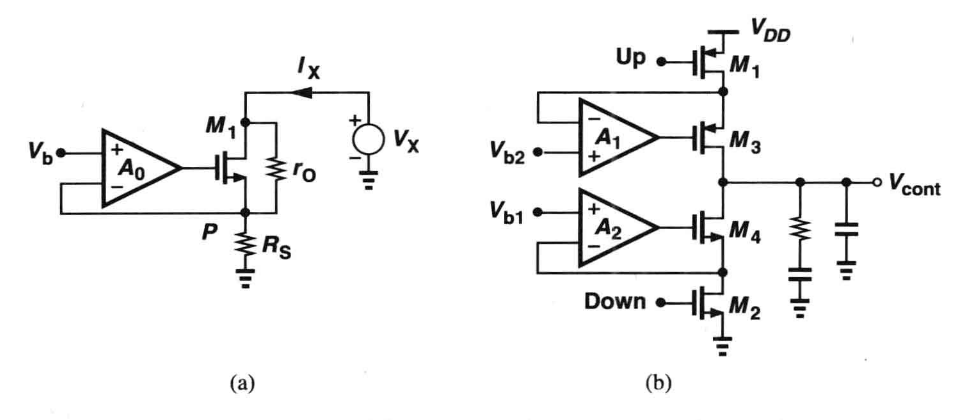


Figure 9.50 *(a) Circuit using an amplifier to raise the output impedance, (b) use of technique in (a) in a charge pump.*

impedance rises. The reader can use a small-signal model to prove that

$$\frac{V_X}{I_X} = (1 + A_0)g_m r_O R_S + r_O + R_S. \tag{9.49}$$

This technique is attractive because it raises the output impedance without consuming additional voltage headroom.

Figure 9.50(b) shows a charge pump employing regulated cascodes. Note that the switches are placed in series with the sources of M_3 and M_4. If the gain of the auxiliary amplifiers is sufficiently large, the mismatch between the Up and Down currents remains small even if M_3 and M_4 enter the triode region by a small amount.

The principal drawback of this approach stems from the finite response of the auxiliary amplifiers. When M_1 and M_2 turn off, the feedback loops around M_3 and M_4 are broken, allowing the outputs of A_1 and A_2 to approach the supply rails. In the next phase comparison instant, these outputs must return and settle to their proper values—a transient substantially longer than the width of the Up and Down pulses ($\approx$ five gate delays). In other words, A_1 and A_2 may simply not have enough time to settle and boost the output impedance according to Eq. (9.49).

Figure 9.51 depicts another technique [7]. Here, M_1-M_4 constitute the main charge pump and M_5-M_8 a replica branch. Note that the bias current of M_{REF} is copied onto M_6 and M_4, and additional transistors M_9 and M_8 imitate the role of M_2 (when it is on). Neglecting random mismatches and assuming a ratio of unity between the CP branch and its replica, we show that the Up current is forced to become equal to the Down current even in the presence of channel-length modulation. We first recognize that, in the locked state, the loop filter serves as a heavy "reservoir," keeping V_{cont} relatively constant. Thus, the servo amplifier, A_0, adjusts the gate voltage of M_5 so as to bring V_X close to V_{cont}. This in turn means that $I_{D6} \approx I_{D4}$ (because $V_{D6} \approx V_{D4}$) and $I_{D5} \approx I_{D3}$ even if the transistors suffer from heavy channel-length modulation. Moreover, since $|I_{D5}| = |I_{D6}|$, we have $|I_{D3}| = I_{D4}$, i.e., the Up and Down currents are equal. The circuit can therefore tolerate a wide output voltage range so long as the open-loop gain of A_0 is sufficiently large to guarantee $V_X \approx V_{cont}$.

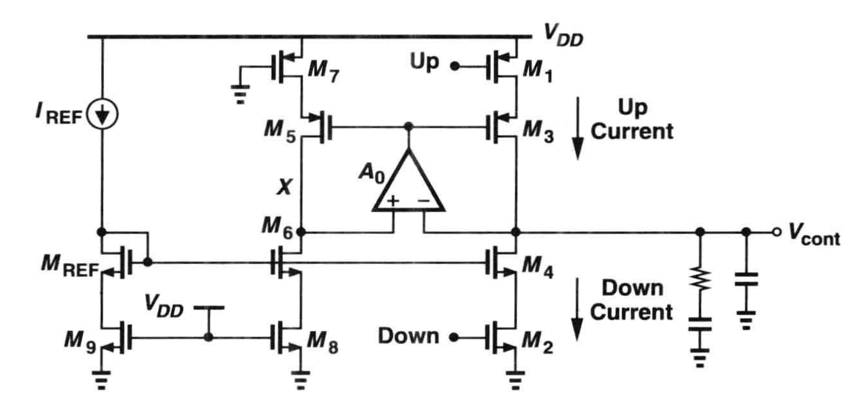


Figure 9.51 *Use of a servo loop to suppress the effect of channel-length modulation.*

A key advantage of this topology over the charge pump in Fig. 9.50(b) is that A_0 need not provide a fast response. This is because, when M_1 and M_2 turn off, the feedback loop consisting of A_0 and M_5 remains intact, thus experiencing a negligible transient.

The performance of the circuit is still limited by random mismatches between the NMOS current sources and between the PMOS current sources. Also, the op amp must operate properly with a nearly rail-to-rail input common-mode range because V_{cont} must come as close to the rails as possible.

Example 9.24

The circuit of Fig. 9.51 contains another feedback loop consisting of A_0 and M_3. In other words, one of the two loops must inevitably have positive feedback. Explain how the feedback polarities are chosen.

Solution:

Since the filter heavily loads the output node, the latter loop is much less agile than the former. We therefore select negative feedback around A_0 and M_5 and positive feedback around A_0 and M_3.

Figure 9.52(a) shows another example using a servo amplifier [8]. In a manner similar to Fig. 9.51, A_0 forces V_X close to V_{cont} such that $I_{D5} \approx I_{D4}$ and $I_{D6} \approx I_{D3}$ (in the absence of random mismatches). Consequently, $|I_{D3}| \approx I_{D4}$. This circuit, however, controls the Up and Down currents through the *gates* of M_3 and M_4, respectively, thereby saving the voltage headroom associated with M_1 and M_2 in Fig. 9.51. This approach is called "gate switching."

The gate switching operation nonetheless exacerbates the problem of Up and Down arrival mismatch. To understand this issue, let us consider the realization shown in Fig. 9.52(b), where the Up and Down pulses have a finite risetime and falltime. We observe that M_1 turns on or off as the Up pulse reaches $V_{DD} - |V_{GS3}| - |V_{TH1}|$, whereas M_2 turns

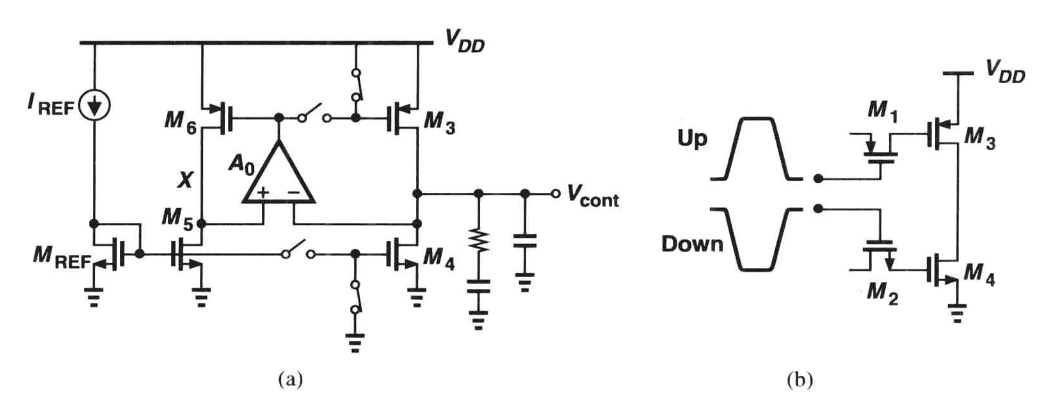


Figure 9.52 *(a) Use of a servo loop to suppress the effect of channel-length modulation in a gate-switched CP, (b) effect of process variations on Up and Down skew.*

on or off as the *Down* pulse crosses $V_{GS4} + V_{TH2}$. Since both of these values vary with process and temperature, it is difficult to ensure that the Up and Down currents arrive simultaneously. Also, op amp A_0 must operate with a wide input voltage range.

Depicted in Fig. 9.53 is another example that cancels both random and deterministic mismatches between the Up and Down currents [9]. In addition to the main output branch consisting of I_1, M_1, M_2, and I_2, the circuit incorporates switches M_5 and M_6, an integrating capacitor, C_X, and an op amp, A_0. Driven by $\overline{\text{Up}}$ and $\overline{\text{Down}}$, the additional switches create a path from I_1 to I_2 when no phase comparison is made. Thus, the *difference* between I_1 and I_2 flows through C_X, monotonically raising or lowering V_X in consecutive input cycles. Op amp A_0 compares this voltage with average V_{cont} and adjusts the value of I_2 so as to bring V_X close to V_{cont}. In other words, in the steady state, V_X remains *constant*, and hence $I_1 = I_2$. The accuracy of the circuit is ultimately limited by the charge injection and clock feedthrough mismatch between M_1 and M_5 and between M_2 and M_6.

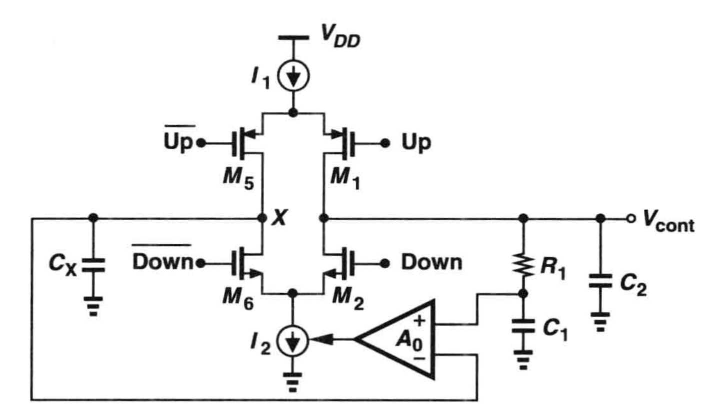


Figure 9.53 *Servo loop around a CP for removing random and deterministic mismatches.*

Example 9.25

A PLL having a reference frequency of f_{REF} and a divide ratio of N exhibits reference sidebands at the output that are 60 dB below the carrier. If the reference frequency is doubled and the divide ratio is halved (so that the output frequency is unchanged), what happens to the reference sidebands? Assume the CP nonidealities are constant and the time during which the CP is on remains much shorter than $T_{REF} = 1/f_{REF}$.

Solution:

Figure 9.54(a) plots the time-domain and frequency-domain behavior of the control voltage in the first case. Since $\Delta T \ll T_{REF}$, we approximate each occurrence of the ripple by an impulse of height $V_0 \cdot \Delta T$. The spectrum of the ripple thus comprises impulses of height $V_0 \cdot \Delta T/T_{REF}$ at harmonics of f_{REF}. The two impulses at $\pm f_{REF}$ can be viewed in the time domain as a sinusoid having a peak amplitude of $2V_0 \cdot \Delta T/T_{REF}$, producing output sidebands that are below the carrier by a factor of $(1/2)(2V_0 \cdot \Delta T/T_{REF})K_{VCO}/(2\pi f_{REF}) = (V_0 \cdot \Delta T K_{VCO})/(2\pi)$.

(*Continues*)

Example 9.25 *(Continued)*

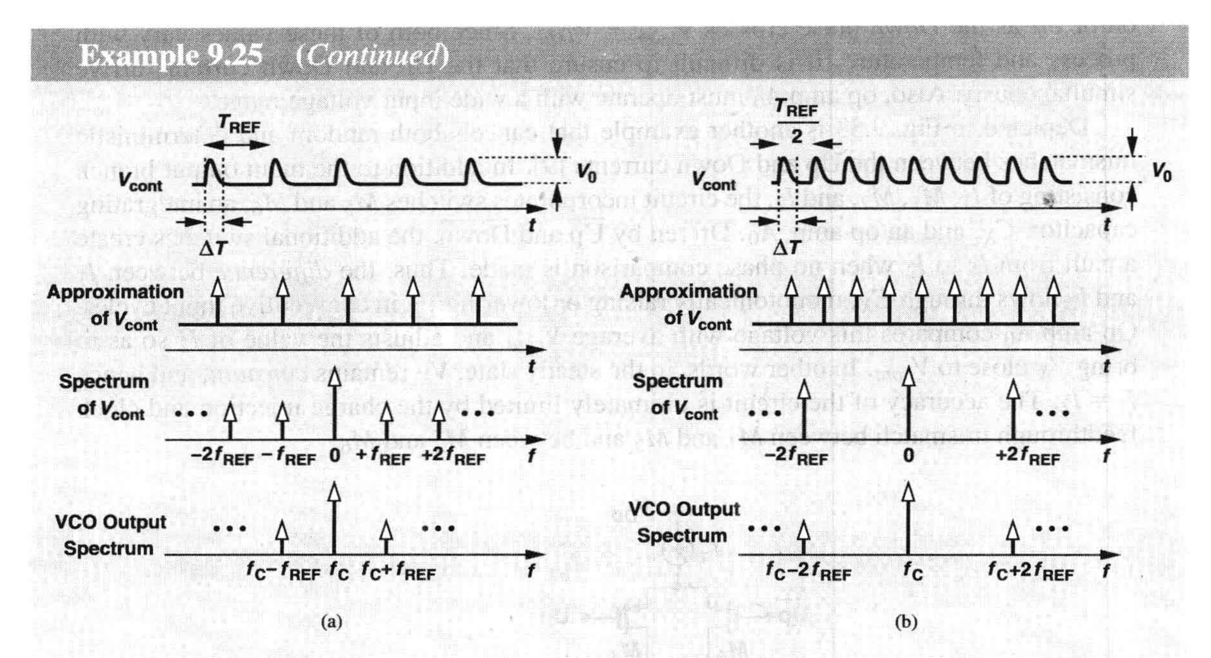


Figure 9.54 *PLL waveforms and output spectrum for an input period of (a) T_{REF}, (b) $T_{REF}/2$.*

Now, consider the second case, shown in Fig. 9.54(b). The ripple repetition rate is doubled, and so is the height of the impulses in the frequency domain. The magnitude of the output sidebands with respect to the carrier is therefore equal to $(1/2)(4V_0 \cdot \Delta T/T_{REF})K_{VCO}/(4\pi f_{REF}) = (V_0 \cdot \Delta T K_{VCO})/(2\pi)$. In other words, the sidebands move away from the carrier but their relative magnitude does not change.

9.5 PHASE NOISE IN PLLS

In our study of oscillators in Chapter 8, we analyzed the mechanism by which device noise translates to phase noise. When an oscillator is phase-locked, its output phase noise profile changes. Also, the reference input to the PLL contains phase noise, corrupting the output. We investigate these effects for type-II PLLs.

9.5.1 VCO Phase Noise

Our understanding of phase-locking suggests that a PLL continually attempts to make the output phase track the input phase. Thus, if the reference input has no phase noise, the PLL attempts to reduce the output phase noise to zero even if the VCO exhibits its own phase noise. From another perspective, as the VCO phase noise accumulates to an appreciable phase error, the loop detects this error and commands the charge pump to briefly turn on and correct it. (If the VCO experienced no phase drift, it would continue to operate at a certain frequency and phase even if the loop were disabled.)

In order to formulate the PLL output noise due to the VCO phase noise, we first derive the transfer function from the VCO phase to the PLL output phase. To this end, we construct the linear phase model of Fig. 9.55(a), where the excess phase of the input is set to zero to

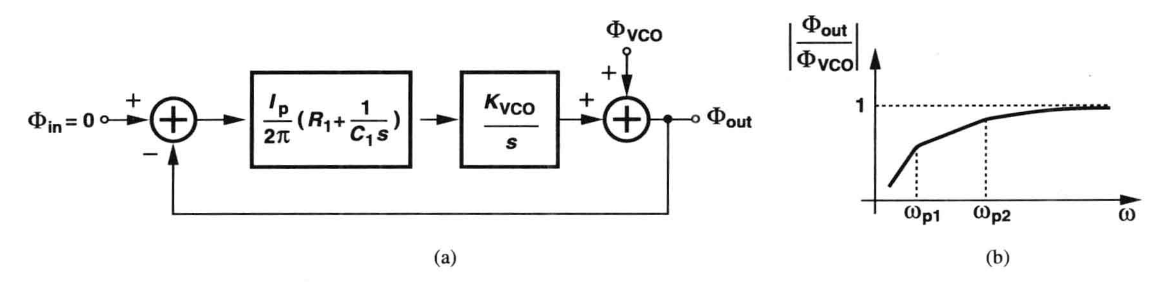


Figure 9.55 *(a) Phase-domain model for studying the effect of VCO phase noise, (b) resulting high-pass response.*

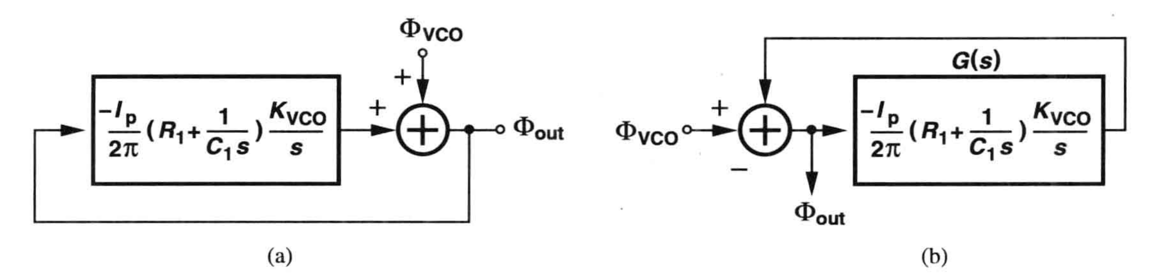


Figure 9.56 *Alternative drawings of phase-domain model showing the effect of VCO phase noise.*

signify a "clean" reference. Beginning from the output, we have

$$-\phi_{out}\left[\frac{I_P}{2\pi}\left(R_1+\frac{1}{C_1 s}\right)\right]\cdot\frac{K_{VCO}}{s}+\phi_{VCO}=\phi_{out}. \tag{9.50}$$

Using the ζ and ω_n expressions developed in Section 9.3.3, we obtain

$$\frac{\phi_{out}}{\phi_{VCO}}=\frac{s^2}{s^2+2\zeta\omega_n s+\omega_n^2}. \tag{9.51}$$

As expected, this transfer function has the same poles as Eq. (9.19), but it also contains two zeros at the origin, exhibiting a *high-pass* behavior [Fig. 9.55(b)].

This result indicates that the PLL suppresses *slow* variations in the phase of the VCO [small ω in Fig. 9.55(b)] but cannot provide much correction for fast variations. In lock, the VCO phase is compared against the input phase and the corresponding error is converted to current, injected into the loop filter to generate a voltage, and finally applied to the VCO so as to counteract its phase variation. Since both the charge pump and the VCO have nearly infinite gain for slowly-varying signals, the negative feedback remains strong for slow phase variations. For fast variations, on the other hand, the loop gain falls and the feedback provides less correction.

From another perspective, the system of Fig. 9.55(a) can be redrawn as shown in Fig. 9.56(a) and hence Fig. 9.56(b). The system $G(s)$ is equivalent to a cascade of two ideal integrators, thus creating a "virtual ground" at its input (at ϕ_{out}). If ϕ_{VCO} varies slowly, ϕ_{out} is near zero, but as ϕ_{VCO} varies faster, $|G(s)|$ falls and the virtual ground experiences larger swings.

Example 9.26

What happens to the frequency response shown in Fig. 9.55(b) if ω_n is increased by a factor of K while ζ remains constant?

Solution:

From Eq. (9.22), we observe that both poles scale up by a factor of K. Since $\phi_{out}/\phi_{VCO} \approx s^2/\omega_n^2$ for $s \approx 0$, the plot is shifted down by a factor of K^2 at low values of ω. Depicted in Fig. 9.57, the response now suppresses the VCO phase noise to a greater extent.

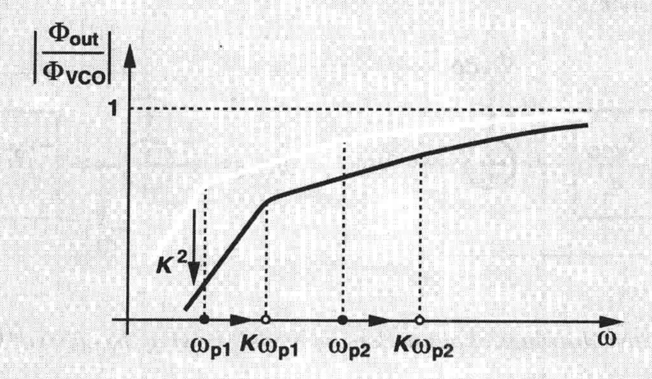


Figure 9.57 *Effect of scaling ω_n by a factor of K on shaped VCO phase noise.*

Example 9.27

Consider a PLL with a feedback divide ratio of N. Compare the phase noise behavior of this case with that of a dividerless loop. Assume the output frequency is unchanged.

Solution:

Redrawing the loop of Fig. 9.56(a) as shown in Fig. 9.58(a), we recognize that the feedback is now weaker by a factor of N. The transfer function given by Eq. (9.51) still applies, but both ζ and ω_n are reduced by a factor of $\sqrt{N}$.

What happens to the magnitude plot of Fig. 9.55(b)? We make two observations. (1) To maintain the same transient behavior, ζ must be constant; e.g., the charge pump current must be scaled up by a factor of N. Thus, the poles given by Eq. (9.22) simply decrease by a factor of $\sqrt{N}$. (2) For $s \to 0$, $\phi_{out}/\phi_{VCO} \approx s^2/\omega_n^2$, which is a factor of N *higher* than that of the dividerless loop. The magnitude of the transfer function thus appears as depicted in Fig. 9.58(b).

A time-domain perspective can also explain the rise in the output phase noise. Assuming that the output frequency remains unchanged in the two cases, we note that the dividerless loop makes phase comparisons—and hence phase corrections—N times more often than the loop with a divider does. That is, in the presence of a divider, the VCO can accumulate phase noise for N cycles without receiving any correction. Figure 9.59 illustrates the two scenarios.

Example 9.27 *(Continued)*

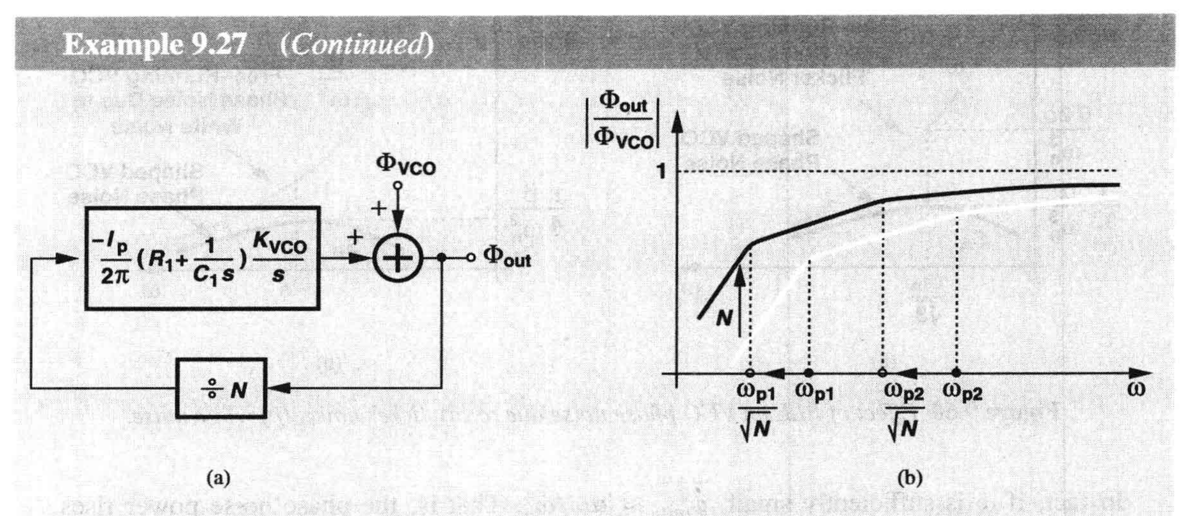


Figure 9.58 *(a) Phase-domain model, and (b) shaped VCO phase noise in the presence of a feedback divider.*

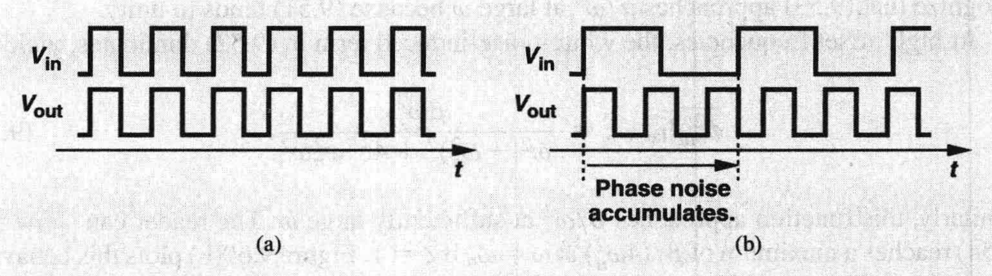


Figure 9.59 *Time-domain PLL waveforms for (a) equal input and output frequencies, (b) input frequency a factor of N lower.*

The PLL output phase noise due to the VCO is equal to the magnitude squared of Eq. (9.51) multiplied by the VCO phase noise. As observed in Chapter 8, oscillator phase noise can be expressed as $(\alpha/\omega^3 + \beta/\omega^2)$, where α and β encapsulate various factors such as the noise injected by devices and the Q, and ω is our notation for the offset frequency ($\Delta\omega$ in Chapter 8). Thus,

$$\overline{\phi_{out}^2} = \frac{\omega^4}{(\omega^2 - \omega_n^2)^2 + 4\zeta^2\omega_n^2\omega^2}\left(\frac{\alpha}{\omega^3} + \frac{\beta}{\omega^2}\right). \tag{9.52}$$

We say the VCO phase noise is "shaped" by the transfer function.

It is instructive to study the above phase noise behavior for low and high offset frequencies. At low offset frequencies (slow VCO phase variations), the flicker-noise-induced term is dominant:

$$\overline{\phi_{out}^2}|_{\text{small }\omega} \approx \frac{\alpha\omega}{(\omega^2 - \omega_n^2)^2 + 4\zeta^2\omega_n^2\omega^2}. \tag{9.53}$$

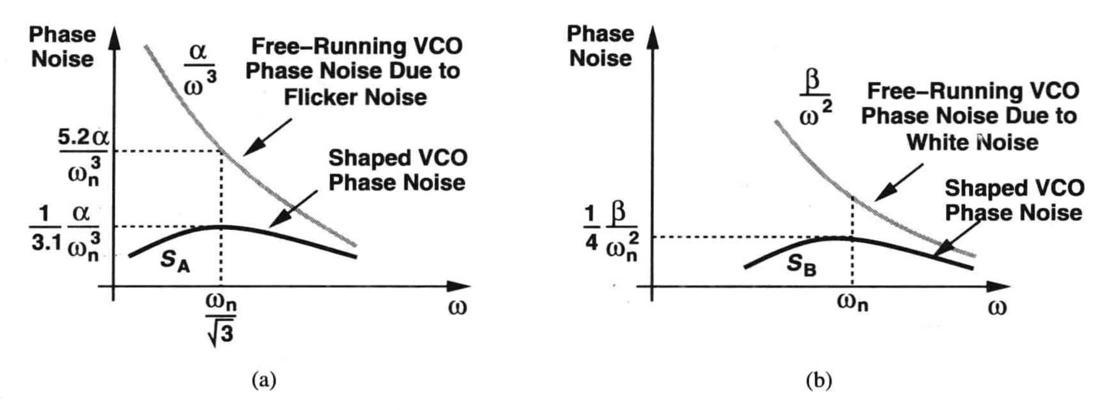


Figure 9.60 *Effect of PLL on VCO phase noise due to (a) flicker noise, (b) white noise.*

In fact, if ω is sufficiently small, $\overline{\phi_{out}^2} \approx \alpha\omega/\omega_n^4$. That is, the phase noise power rises linearly with frequency. The reader can show that Eq. (9.53) reaches a maximum of $9\alpha/(16\sqrt{3}\omega_n^3) \approx \alpha/(3.1\omega_n^3)$ at $\omega = \omega_n/\sqrt{3}$ if $\zeta = 1$. Figure 9.60(a) plots this behavior, indicating that the phase-locked VCO exhibits 12 dB less phase noise at $\omega_n/\sqrt{3}$. We recognize that (9.53) approaches α/ω^3 at large ω because (9.51) tends to unity.

At high offset frequencies, the white-noise-induced term in (9.52) dominates, yielding

$$\overline{\phi_{out}^2}|_{\text{large }\omega} = \frac{\beta\omega^2}{(\omega^2 - \omega_n^2)^2 + 4\zeta^2\omega_n^2\omega^2}. \tag{9.54}$$

Similarly, this function approaches β/ω^2 at sufficiently large ω. The reader can show that (9.54) reaches a maximum of $\beta/(4\omega_n^2)$ at $\omega = \omega_n$ if $\zeta = 1$. Figure 9.60(b) plots this behavior, suggesting a 6-dB reduction at ω_n. In practice, the overall output phase noise is a combination of these two results.

Figure 9.61 summarizes our findings. In addition to the free-running VCO phase noise, the curves corresponding to α/ω^3 and β/ω^2 are also drawn. The overall PLL output phase noise is equal to the sum of S_A and S_B. However, the actual shape depends on two factors: (1) the intersection frequency of α/ω^3 and β/ω^2, and (2) the value of ω_n. The following example illustrates these dependencies.

Example 9.28

Sketch the overall output phase noise if (a) the intersection of α/ω^3 and β/ω^2 lies at a low frequency and ω_n is quite larger than that, (b) the intersection of α/ω^3 and β/ω^2 lies at a high frequency and ω_n is quite smaller than that. (These two cases represent high and low thermal-noise-induced phase noise, respectively.)

Solution:

Depicted in Fig. 9.62(a), the first case contains little $1/f$ noise contribution, exhibiting a shaped phase noise, S_{out}, that merely follows β/ω^2 at large offsets. The second case,

Example 9.28 *(Continued)*

shown in Fig. 9.62(b), is dominated by the shaped $1/f$ noise regime and provides a shaped spectrum nearly equal to the free-running VCO phase noise beyond roughly $\omega = \omega_n$. We observe that the PLL phase noise experiences more peaking in the latter case.

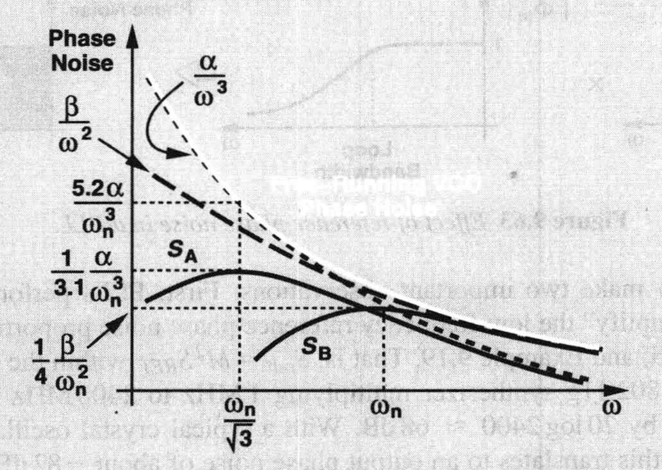


Figure 9.61 *Shaped VCO phase noise summary.*

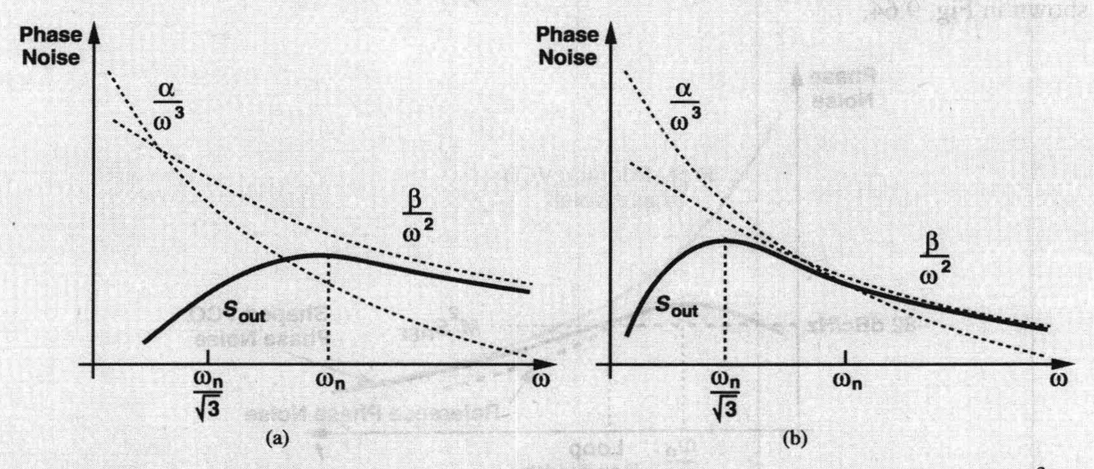


Figure 9.62 *Shaped VCO phase noise for (a) low, and (b) high intersection frequencies of α/ω^3 and β/ω^2.*

9.5.2 Reference Phase Noise

The reference phase noise is simply shaped by the input/output transfer function of the PLL. From Eq. (9.19), we write

$$S_{out} = \frac{4\zeta^2\omega_n^2\omega^2 + \omega_n^4}{(\omega^2 - \omega_n^2)^2 + 4\zeta^2\omega_n^2\omega^2} S_{REF}, \tag{9.55}$$

where S_{REF} denotes the reference phase noise. Note that crystal oscillators providing the reference typically display a *flat* phase noise profile beyond an offset of a few kilohertz. The overall behavior is shown in Fig. 9.63.

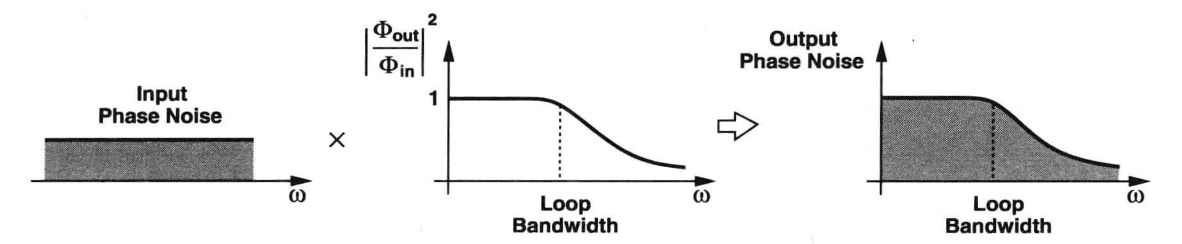


Figure 9.63 *Effect of reference phase noise in a PLL.*

We must now make two important observations. First, PLLs performing frequency multiplication "amplify" the low-frequency reference phase noise proportionally. This can be seen from (9.35) and Example 9.19. That is, $S_{out} = M^2 S_{REF}$ within the loop bandwidth. For example, an 802.11g synthesizer multiplying 1 MHz to 2400 MHz raises the reference phase noise by $20 \log 2400 \approx 68$ dB. With a typical crystal oscillator phase noise of -150 dBc/Hz, this translates to an output phase noise of about -82 dBc/Hz within the loop bandwidth. A loop bandwidth of around 100 kHz therefore results in the output profile shown in Fig. 9.64.

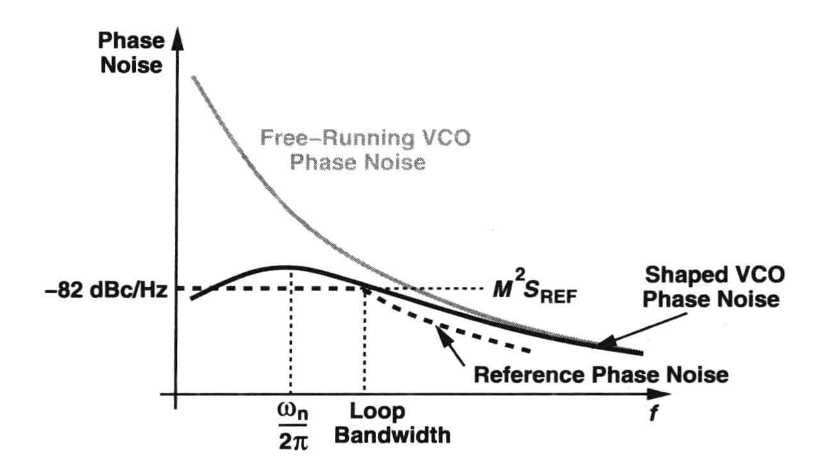


Figure 9.64 *Example of reference and shaped VCO phase noise in a PLL.*

The phase noise multiplication can also be analyzed in the time domain: if the input edges are (slowly) displaced by ΔT seconds ($2\pi \Delta T / T_{REF}$ radians), then the output edges are also displaced by ΔT seconds, which amounts to $2\pi \Delta T / (T_{REF}/N)$ radians and hence $20 \log N$ decibels of higher phase noise.

Second, the total phase noise at the output (the area under the phase noise profile in Fig. 9.63) increases with the loop bandwidth—a trend opposite of that observed for the VCO phase noise. In other words, the choice of the loop bandwidth entails a trade-off between the reference and the VCO phase noise contributions.

9.6 LOOP BANDWIDTH

As seen in this chapter, the bandwidth of the PLL plays a critical role in the overall performance. Our observations thus far indicate that (1) the settling behavior can be roughly characterized by a time constant in the range of $1/(\zeta\omega_n)$ and $1/(2\zeta\omega_n)$ depending on the value of ζ (Example 9.18); (2) the continuous-time approximation requires that the PLL time constant be much longer than the input period; (3) if the PLL bandwidth increases, the VCO phase noise is suppressed more heavily while the reference phase noise appears across a larger bandwidth at the output.

But how should the loop bandwidth be defined? We can simply compute the -3-dB bandwidth by equating the magnitude squared of Eq. (9.19) to $1/2$:

$$\frac{(2\zeta\omega_n\omega_{-3dB})^2 + \omega_n^4}{(\omega_{-3dB}^2 - \omega_n^2)^2 + (2\zeta\omega_n\omega_{-3dB})^2} = \frac{1}{2}. \tag{9.56}$$

It follows that

$$\omega_{-3dB}^2 = [1 + 2\zeta^2 + \sqrt{(1 + 2\zeta^2)^2 + 1}]\omega_n^2. \tag{9.57}$$

For example, if ζ lies in the range of $\sqrt{2}/2$ and 1, then ω_{-3dB} is between $2.1\omega_n$ and $2.5\omega_n$. Also, if $2\zeta^2 \gg 1$, then $\omega_{-3dB} \approx 2\zeta\omega_n$, as predicted by the one-pole approximation in Section 9.3.4. Figure 9.65(a) plots $|\phi_{out}/\phi_{in}|$ and $|\phi_{out}/\phi_{VCO}|$ for $\zeta = 1$. Also shown is the shaped VCO phase noise for the white noise regime. Figure 9.65(b) repeats these plots for the case of $\zeta^2 \gg 1$.

In the design of PLLs, we impose a loop time constant much longer than the input period, T_{in}, or a loop bandwidth much smaller than the input frequency to ensure a well-behaved settling. These two constraints, however, are not exactly equivalent. For example,

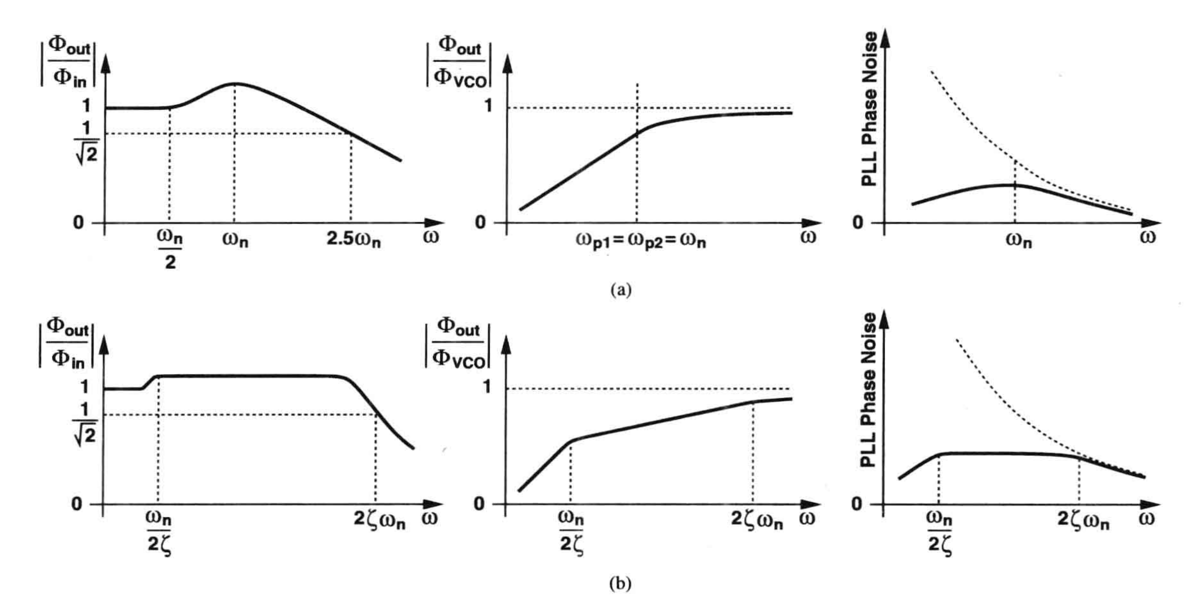


Figure 9.65 *Frequency responses and shaped VCO phase noise for (a) $\zeta = 1$, and (b) $\zeta^2 \gg 1$.*

if ζ is around unity, the former translates to

$$\frac{1}{\zeta\omega_n} \gg T_{in} \tag{9.58}$$

whereas the latter yields

$$2.5\omega_n \ll \omega_{in}. \tag{9.59}$$

Equation (9.59) is a stronger condition and is usually enforced. For higher values of ζ, the loop bandwidth approaches $2\zeta\omega_n$ and is set to approximately one-tenth of ω_{in}.

9.7 DESIGN PROCEDURE

The design of PLLs begins with the building blocks: the VCO is designed according to the criteria and the procedure described in Chapter 8; the feedback divider is designed to provide the required divide ratio and operate at the maximum VCO frequency (Chapter 10); the PFD is designed with careful attention to the matching of the Up and Down pulses; and the charge pump is designed for a wide output voltage range, minimal channel-length modulation, etc. In the next step, a loop filter must be chosen and the building blocks must be assembled so as to form the PLL.

In order to arrive at a well-behaved PLL design, we must properly select the charge pump current and the loop filter components. We begin with two governing equations:

$$\zeta = \frac{R_1}{2}\sqrt{\frac{I_p C_1 K_{VCO}}{2\pi M}} \tag{9.60}$$

$$\omega_n = \sqrt{\frac{I_p K_{VCO}}{2\pi C_1 M}}, \tag{9.61}$$

and choose

$$\zeta = 1 \tag{9.62}$$

$$2.5\omega_n = \frac{1}{10}\omega_{in}. \tag{9.63}$$

Since K_{VCO} is known from the design of the VCO, we now have two equations and three unknowns, namely, I_p, C_1, and R_1; i.e., the solution is not unique. In particular, the charge pump current can be chosen in the range of a few tens of microamperes to a few milliamperes. With I_p selected, C_1 is obtained from (9.61) and (9.63), and R_1 from (9.60). Lastly, we choose the second capacitor (C_2 in Fig. 9.38) to be about $0.2C_1$. We apply this procedure to the design of a synthesizer in Chapter 13.

Example 9.29

A PLL must generate an output frequency of 2.4 GHz from a 1-MHz reference. If $K_{VCO} = 300$ MHz/V, determine the other loop parameters.

Solution:

We select $\zeta = 1$, $2.5\omega_n = \omega_{in}/10$, i.e., $\omega_n = 2\pi(40 \text{ kHz})$, and $I_p = 500\,\mu\text{A}$. Substituting $K_{VCO} = 2\pi \times (300\,\text{MHz/V})$ in (9.61) yields $C_1 = 0.99$ nF. This large value necessitates an off-chip capacitor. Next, (9.60) gives $R_1 = 8.04\,\text{k}\Omega$. Also, $C_2 = 0.2$ nF. As explained in Appendix I, the choice of $\zeta = 1$ and $C_2 = 0.2C_1$ automatically guarantees the condition expressed by (9.41).

Since C_1 is quite large, we can revise our choice of I_p. For example, if $I_p = 100\,\mu\text{A}$, then (9.61) yields $C_1 = 0.2$ nF (still quite large). But, for $\zeta = 1$, R_1 must be raised by a factor of 5, i.e., $R_1 = 40.2\,\text{k}\Omega$. Also, $C_2 = 40$ pF.

9.8 APPENDIX I: PHASE MARGIN OF TYPE-II PLLS

In this appendix, we derive the phase margin of second-order and third-order type-II PLLs. Consider the open-loop magnitude and phase response of a second-order PLL as shown in Fig. 9.66. The magnitude falls with a slope of -40 dB/dec up to the zero frequency, $\omega_z = (R_1C_1)^{-1}$, at which the slope changes to -20 dB/dec. The phase begins at -180° and reaches -135° at the zero frequency. To determine the phase margin, we must compute the phase contribution of the zero at the unity-gain frequency, ω_u. Let us first calculate the value of ω_u.

$$\left|\frac{I_p}{2\pi}\left(R_1 + \frac{1}{C_1 s}\right)\frac{K_{VCO}}{s}\right|^2_{s=j\omega_u} = 1 \tag{9.64}$$

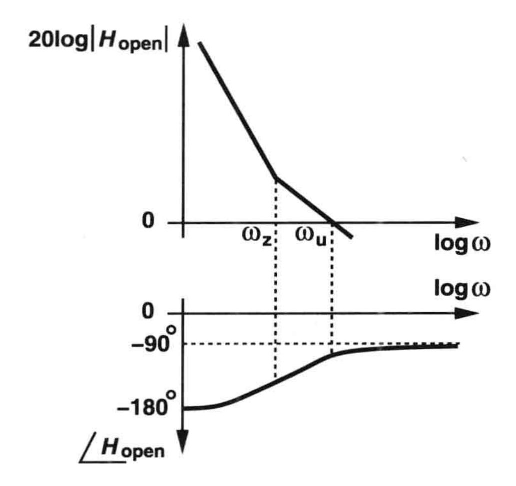


Figure 9.66 *Open-loop magnitude and phase response of type-II PLL.*

and hence

$$\left(\frac{I_p K_{VCO}}{2\pi}\right)^2 \frac{R_1^2 C_1^2 \omega_u^2 + 1}{C_1^2 \omega_u^4} = 1. \tag{9.65}$$

Using Eqs. (9.20) and (9.21) as short-hand notations and noting that $R_1 C_1 \omega_n^2 = 2\zeta\omega_n$, we have

$$-\omega_u^4 + 4\zeta^2 \omega_n^2 \omega_u^2 + \omega_n^4 = 0, \tag{9.66}$$

and

$$\omega_u^2 = \left(2\zeta^2 + \sqrt{4\zeta^4 + 1}\right)\omega_n^2. \tag{9.67}$$

The phase margin is therefore given by

$$PM = \tan^{-1}\frac{\omega_u}{\omega_z} = \tan^{-1} R_1 C_1 \omega_u \tag{9.68}$$

$$= \tan^{-1}\left(2\zeta\sqrt{2\zeta^2 + \sqrt{4\zeta^4 + 1}}\right). \tag{9.69}$$

For example, if $\zeta = 1$, then $PM = 76^\circ$ and $\omega_u/\omega_z \approx 4$, and if $\zeta = \sqrt{2}/2$, then $PM = 65^\circ$ and $\omega_u/\omega_z \approx 2.2$. For $\zeta \geq \sqrt{2}/2$, we have $\sqrt{4\zeta^4 + 1} \approx 2\zeta^2 + 1/(4\zeta^2)$ and hence

$$PM \approx \tan^{-1}\left(2\zeta\sqrt{4\zeta^2 + \frac{1}{4\zeta^2}}\right) \tag{9.70}$$

$$\approx \tan^{-1}\left[4\zeta^2\left(1 + \frac{1}{32\zeta^4}\right)\right]. \tag{9.71}$$

Example 9.30

Sketch the open-loop characteristics of the PLL with R_1 or C_1 as a variable.

Solution:

As R_1 increases, ω_z falls but ω_u rises (because the slope of $|H_{open}|$ must still be equal to −20 dB/dec) [Fig. 9.67(a)]. On the other hand, as C_1 increases, ω_z falls and ω_u remains relatively constant [Fig. 9.67(b)]. This is because, for $\zeta \geq \sqrt{2}/2$,

$$\omega_u \approx \sqrt{4\zeta^2 + \frac{1}{4\zeta^2}}\omega_n \tag{9.72}$$

$$\approx 2\zeta\left(1 + \frac{1}{32\zeta^4}\right)\omega_n, \tag{9.73}$$

Example 9.30 *(Continued)*

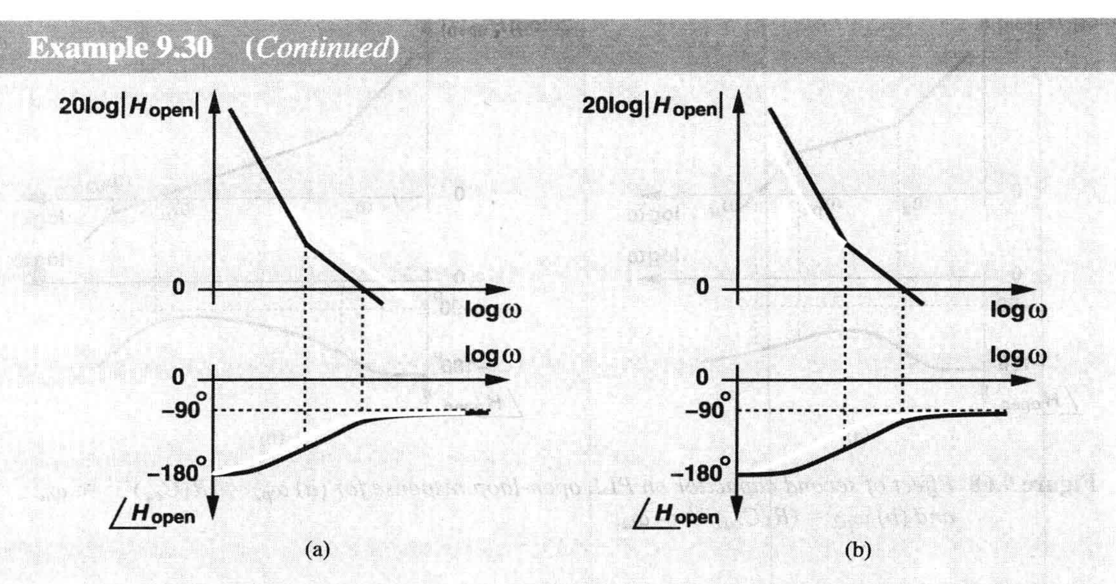


Figure 9.67 *Effect of (a) higher R_1, or (b) higher C_1 on frequency response of type-II PLL.*

which, in most cases of practical interest, can be written as

$$\omega_u \approx 2\zeta\omega_n \tag{9.74}$$

$$\approx \frac{R_1 I_p K_{VCO}}{2\pi}. \tag{9.75}$$

The two trends depicted in Figs. 9.67(a) and (b) also shed light on the stronger dependence of ζ on R_1 than on C_1: in the former, the PM increases because ω_z falls *and* ω_u rises, whereas in the latter, the PM increases only because ω_z falls.

Let us now consider the third-order loop of Fig. 9.38. The reader can show that the PFD/CP/filter cascade provides the following transfer function:

$$\frac{V_{cont}}{\Delta\phi}(s) = \frac{I_p}{2\pi}\cdot\frac{R_1C_1s+1}{R_1C_{eq}s+1}\cdot\frac{1}{(C_1+C_2)s}, \tag{9.76}$$

where $C_{eq} = C_1C_2/(C_1 + C_2)$. The pole contributed by the filter, ω_{p2}, thus lies at $-(R_1C_{eq})^{-1}$. Figure 9.68(a) plots an example of the open-loop frequency response, revealing the PM degradation due to C_2.

How should ω_{p2} be chosen? If located *below* ω_u, this pole yields a PM of *less* than 45°. This is because the phase profile shown in Fig. 9.68(a) experiences a contribution of -45° from ω_{p2} at ω_{p2} and hence a more negative amount at ω_u. For this reason, ω_{p2} must be chosen *higher* than ω_u [Fig. 9.68(b)]. The key point here is that the magnitude of ω_u is roughly the same even in the presence of ω_{p2}, allowing the use of Eq. (9.73).

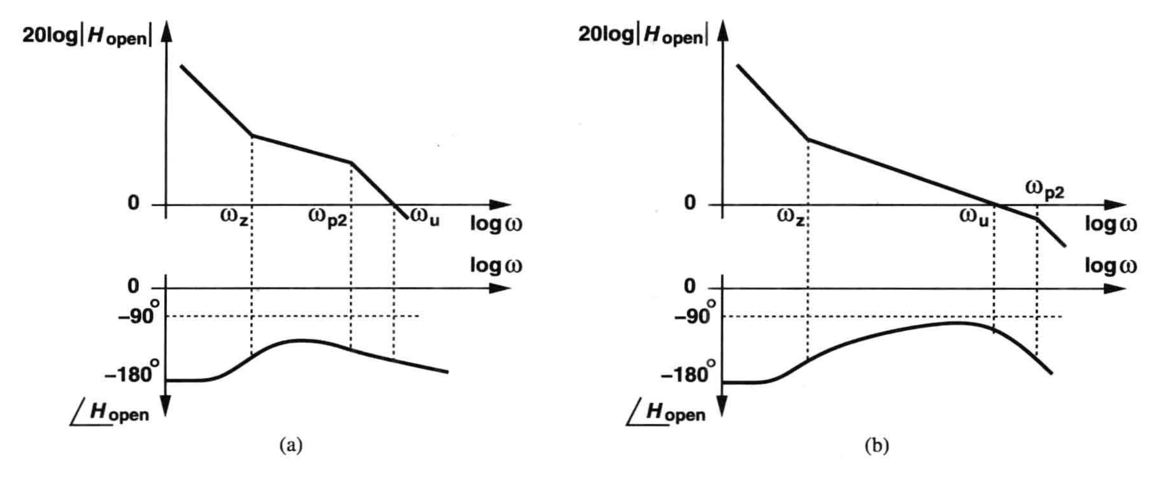


Figure 9.68 *Effect of second capacitor on PLL open-loop response for (a) $\omega_{p2} = (R_1 C_{eq})^{-1} < \omega_u$, and (b) $\omega_{p2} = (R_1 C_{eq})^{-1} > \omega_u$.*

The phase margin can now be calculated as

$$\text{PM} = \tan^{-1}\frac{\omega_u}{\omega_z} - \tan^{-1}\frac{\omega_u}{\omega_{p2}} \tag{9.77}$$

$$= \tan^{-1}(R_1 C_1 \omega_u) - \tan^{-1}(R_1 C_{eq} \omega_u) \tag{9.78}$$

$$= \tan^{-1}\left[4\zeta^2\left(1 + \frac{1}{32\zeta^2}\right)\right] - \tan^{-1}\left[4\zeta^2\frac{C_{eq}}{C_1}\left(1 + \frac{1}{32\zeta^2}\right)\right]. \tag{9.79}$$

In most cases of practical interest, $32\zeta^2 \gg 1$ and hence

$$\text{PM} \approx \tan^{-1}(4\zeta^2) - \tan^{-1}\left(4\zeta^2\frac{C_{eq}}{C_1}\right). \tag{9.80}$$

Note that this result is valid only if $\zeta \geq 1$ and ω_{p2} is well above ω_u.

An alternative approach seeks that value of ω_u which maximizes the PM in Eq. (9.78) [10]. Differentiation yields

$$\omega_u = \frac{1}{R_1 C_1}\sqrt{1 + \frac{C_1}{C_2}} \tag{9.81}$$

and

$$\text{PM}_{max} = \tan^{-1}\left(\frac{C_1/C_2}{2\sqrt{1 + C_1/C_2}}\right). \tag{9.82}$$

The corresponding ζ can be obtained by differentiating Eq. (9.80):

$$\zeta = \frac{1}{2}\sqrt[4]{\frac{C_1}{C_{eq}}}, \tag{9.83}$$

approximately equal to 0.783 for $C_1 = 5C_2$.

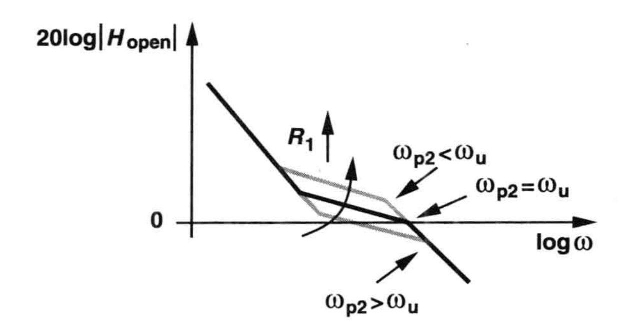


Figure 9.69 *Effect of higher R_1 on PLL frequency response in the presence of second capacitor.*

The foregoing study also reveals another important limitation in the choice of the loop parameters: with C_2 present, R_1 cannot be arbitrarily large. After all, if $R_1 \to \infty$, the series combination of R_1 and C_1 vanishes, leaving only C_2 and hence only two ideal integrators in the loop. To determine an upper bound on R_1, we note that, as R_1 increases, ω_{p2} approaches and eventually falls below ω_u (Fig. 9.69). If we consider $\omega_{p2} \approx \omega_u$ as the lower limit on ω_{p2}, then

$$\frac{1}{R_1 C_{eq}} \geq 2\zeta\omega_n \tag{9.84}$$

$$\geq \frac{R_1 I_p K_{VCO}}{2\pi}. \tag{9.85}$$

It follows that

$$R_1^2 \leq \frac{2\pi}{I_p K_{VCO} C_{eq}}, \tag{9.86}$$

and hence

$$\frac{C_2}{C_1 + C_2} \leq \frac{1}{4\zeta^2}. \tag{9.87}$$

We note that, if $\zeta \approx 1$ and $C_2 \approx 0.2C_1$, this condition is satisfied.

REFERENCES

[1] F. M. Gardner, *Phaselock Techniques*, Second Edition, New York: Wiley & Sons, 1979.

[2] B. Razavi, *Design of Analog CMOS Integrated Circuits,* Boston: McGraw-Hill, 2001.

[3] J. P. Hein and J. W. Scott, "z-Domain Model for Discrete-Time PLLs," *IEEE Trans. Circuits and Systems,* vol. 35, pp. 1393–1400, Nov. 1988.

[4] J. Alvarez et al., "A Wide-Bandwidth Low-Voltage PLL for PowerPC Microprocessors," *IEEE J. of Solid-State Circuits,* vol. 30, pp. 383–391, April 1995.

[5] J. M. Ingino and V. R. von Kaenel, "A 4-GHz Clock System for a High-Performance System-on-a-Chip Design," *IEEE J. of Solid-State Circuits,* vol. 36, pp. 1693–1699, Nov. 2001.

[6] B. J. Hosticka, "Improvement of the Gain of CMOS Amplifiers," *IEEE J. of Solid-State Circuits,* vol. 14, pp. 1111–1114, Dec. 1979.

[7] J.-S. Lee et al., "Charge Pump with Perfect Current Matching Characteristics in Phase-Locked Loops," *Electronics Letters,* vol. 36, pp. 1907–1908, Nov. 2000.

[8] M. Terrovitis et al., "A 3.2 to 4 GHz 0.25 μm CMOS Frequency Synthesizer for IEEE 802.11a/b/g WLAN," *ISSCC Dig. Tech. Papers,* pp. 98–99, Feb. 2004.

[9] M. Wakayam, "Low offset and low glitch energy charge pump and method of operating same," US Patent 7057465, April 2005.

[10] H. R. Rategh, H. Samavati, and T. H. Lee, "A CMOS Frequency Synthesizer with an Injection-Locked Frequency Divider for a 5-GHz Wireless LAN Receiver," *IEEE J. of Solid-State Circuits,* vol. 35, pp. 780–788, May 2000.

PROBLEMS

9.1. The mixer phase detector characteristic shown in Fig. 9.5 exhibits a *zero* gain at the peaks, e.g., at $\Delta\phi = 0$. A PLL using such a PD would therefore suffer from a zero loop gain at these points. Does this mean the PLL would not lock?

9.2. If K_{VCO} in the PLL of Fig. 9.10(a) is very high and the PD has the characteristic shown in Fig. 9.5, can we estimate the value of $\Delta\phi$?

9.3. Repeat Problem 9.2 if the sign of K_{VCO} is changed.

9.4. Determine at what frequencies the output sidebands of Fig. 9.7(a) are located. Are these sidebands or harmonics?

9.5. In the PLL of Fig. 9.8(b), an input change of $\Delta\phi$ exactly yields an output change of $\Delta\phi$. On the other hand, in the buffer of Fig. 9.8(a), an input change of ΔV produces an output change of $\Delta V/(A_0 + 1)$, where A_0 is the open-loop gain of the op amp. How do we explain this difference?

9.6. Suppose the PLL of Fig. 9.12 is locked. Now, we replace R_1 with an open circuit. What happens at the output as time passes? Consider two cases: a noiseless VCO and a noisy VCO. This example shows that if the VCO (excess) phase does not drift with time, the feedback loop can be broken.

9.7. Determine the transfer function, ζ, and ω_n for the frequency-multiplying PLL of Fig. 9.18(b).

9.8. For the PFD of Fig. 9.20, determine whether or not the average value of $Q_A - Q_B$ is a linear function of the input frequency difference.

9.9. Compute the peak value of $|H|$ in Example 9.17.

9.10. Suppose a PLL designed with $\zeta = 1$, a loop bandwidth of $\omega_{in}/25$, and a tuning range of 10%. Assume V_{cont} can vary from 0 to V_{DD}. Prove that that the voltage drop across the loop filter resistor reaches roughly $1.6\pi V_{DD}$ if no second capacitor is used.

9.11. A PLL is designed with an input frequency of 1 MHz and an output frequency of 1 GHz. Now suppose the design is modified to operate with an input frequency of 2 MHz. Explain from Eq. (9.43) what happens to the output sidebands if (a) the output frequency remains unchanged, or (b) the output frequency also doubles. Assume in the latter case that K_{VCO} must double.

9.12. The ripple on the control voltage creates sidebands around the carrier at the output of a PLL, equivalently disturbing the phase of the VCO. Explain why the PLL

suppresses the VCO phase noise (within the loop bandwidth) but not the sidebands due to the ripple.

9.13. Consider the PLL shown in Fig. 9.70, where amplifier A_1 is interposed between the filter and the VCO. If the amplifier exhibits an input-referred flicker noise density given by α/f, determine the PLL output phase noise.

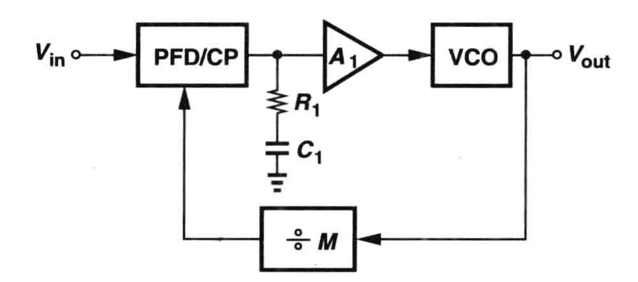


Figure 9.70 *PLL with amplifier in the loop.*

9.14. A PLL incorporates a VCO having the characteristic shown in Fig. 9.71. It is possible to compensate for the VCO nonlinearity by varying the charge pump current as a function of the control voltage so that the loop dynamics remain relatively constant. Sketch the desired variation of the charge pump current.

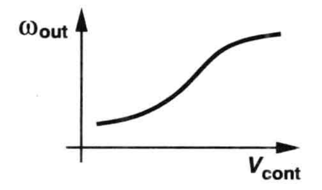


Figure 9.71 *Nonlinear characteristic of a VCO.*

9.15. A PLL operates with input and output frequencies equal to f_1. Suppose the input frequency and hence the output frequency are changed to $f_1/2$. Assuming all loop parameters remain unchanged and neglecting the continuous-time approximation issues, explain which one of these arguments is correct and why the other one is not:

(a) The PFD now makes half as many phase comparisons per second, pumping half as much charge into the loop filter. Thus, the loop is less stable.

(b) Equation $\zeta = (R_P/2)\sqrt{(I_P K_{VCO} C_P)/(2\pi)}$ indicates that ζ remains constant and the loop is as stable as before.

9.16. In the loop shown in Fig. 9.72, V_{ex} suddenly jumps by ΔV. Sketch the waveforms for V_{cont} and V_{LPF} and determine the total change in V_{cont}, V_{LPF}, the output frequency, and the input-output phase difference.

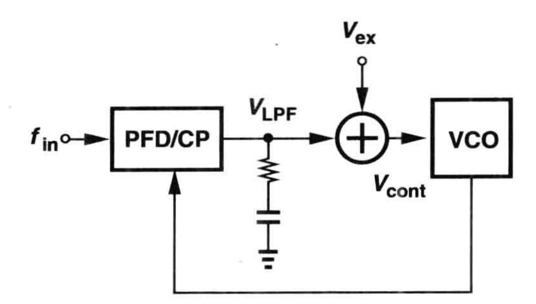


Figure 9.72 *PLL with a step on the control voltage.*

9.17. Two PLL configurations are shown in Fig. 9.73. Assume the SSB mixer *adds* its input frequencies. Also, assume f_1 is a constant frequency provided externally and it is less than f_{REF}. The control voltage experiences a small sinusoidal ripple with a frequency of f_{REF}. Both PLLs are locked.

(a) Determine the output frequencies of the two PLLs.

(b) Determine the spectrum at point A due to the ripple.

(c) Now determine the spectrum at nodes B and C.

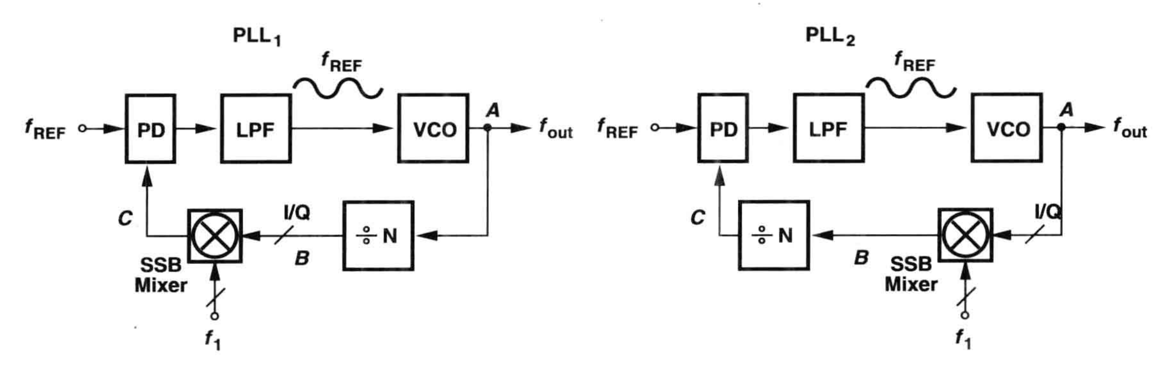
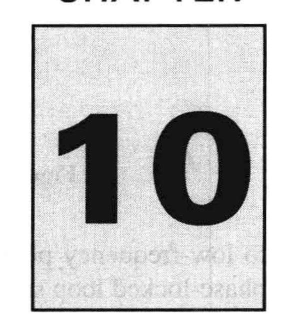


Figure 9.73 *Two PLL topologies.*

CHAPTER



INTEGER-N FREQUENCY SYNTHESIZERS

The oscillators used in RF transceivers are usually embedded in a "synthesizer" environment so as to precisely define their output frequency. Synthesizer design has for decades proved a difficult task, leading to hundreds of RF synthesis techniques. RF synthesizers typically employ phase-locking and must deal with the generic PLL issues described in Chapter 9. In this chapter, we study one class called "integer-N" synthesizers. The chapter outline is shown below. The reader is encouraged to first review Chapters 8 and 9.

Basic Synthesizer	PLL-Based Modulation	Divider Design
Settling Behavior	In-Loop Modulation	Pulse-Swallow Divider
Spur Reduction Techniques	Offset-PLL TX	Dual-Modulus Dividers
		CML and TSPC Techniques
		Miller and Injection-Locked Dividers

10.1 GENERAL CONSIDERATIONS

Recall from Chapter 3 that each wireless standard provides a certain number of frequency channels for communication. For example, Bluetooth has 80 1-MHz channels in the range of 2.400 GHz to 2.480 GHz. At the beginning of each communication session, one of these channels, f_j, is allocated to the user, requiring that the LO frequency be set (defined) accordingly (Fig. 10.1). The synthesizer performs this precise setting.

The reader may wonder why a synthesizer is necessary. It appears that the control voltage of a VCO can be simply changed to establish the required LO frequency. The VCOs studied in Chapter 8 are called "free-running" because, for a given control voltage, their output frequency is defined by the circuit and device parameters. The frequency therefore varies with temperature, process, and supply voltage. It also drifts with time due

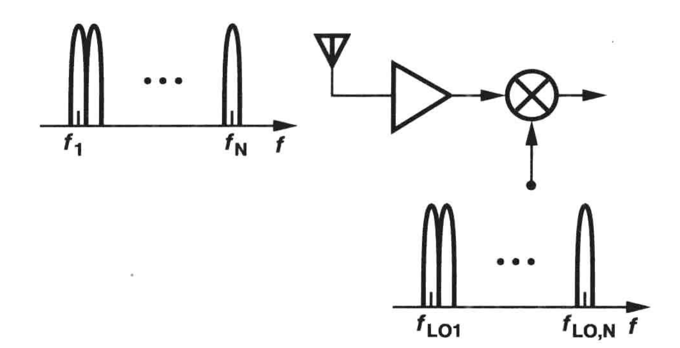


Figure 10.1 *Setting of LO frequency for each received channel.*

to low-frequency phase noise components. For these reasons, VCOs are controlled by a phase-locked loop such that their output frequency can track a precise reference frequency (typically derived from a crystal oscillator).

The high precision expected of the LO frequency should not come as a surprise. After all, narrow, tightly-spaced channels in wireless standards tolerate little error in transmit and receive carrier frequencies. For example, as shown in Fig. 10.2, a slight shift leads to significant spillage of a high-power interferer into a desired channel. We know from the wireless standards studied in Chapter 3 that the channel spacing can be as small as 30 kHz while the center frequency lies in the gigahertz range.

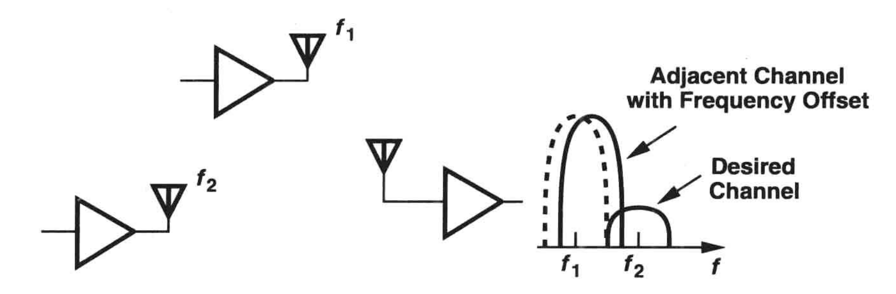


Figure 10.2 *Effect of LO frequency error in TX.*

Figure 10.3 shows the conceptual picture we have thus far formed of a synthesizer. The output frequency is generated as a multiple of a precise reference, f_{REF}, and this multiple is changed by the channel selection command so as to cover the carrier frequencies required by the standard.

In addition to accuracy and channel spacing, several other aspects of synthesizers impact a transceiver's performance: phase noise, sidebands, and "lock time." We studied the effect of phase noise in Chapter 8 and deal with sidebands and lock time here. We know that, if the control voltage of a VCO is periodically disturbed, then the output spectrum contains sidebands symmetrically disposed around the carrier. This indeed occurs if a VCO is placed in a phase-locked loop and experiences the ripple produced by the PFD and

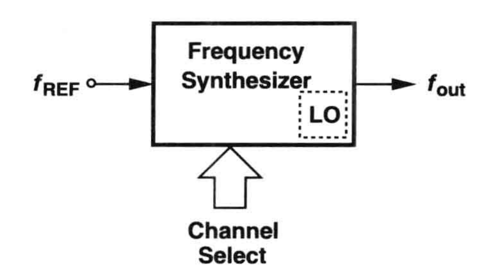


Figure 10.3 *Generic frequency synthesizer.*

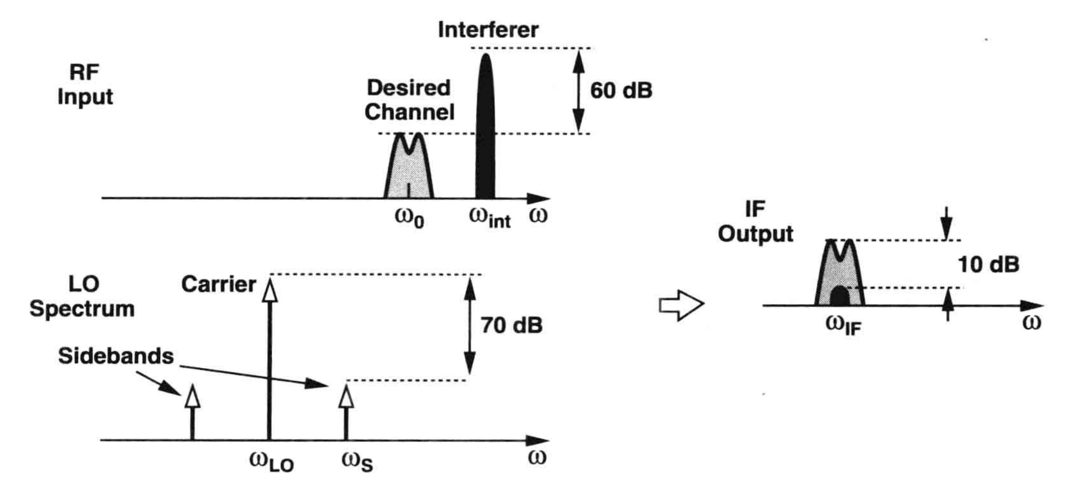


Figure 10.4 *Reciprocal mixing.*

CP nonidealities. We therefore wish to understand the effect of such sidebands ("spurs"). Illustrated in Fig. 10.4, the effect of sidebands is particularly troublesome in the receiver path. Suppose the synthesizer (the LO) output consists of a carrier at ω_{LO} and a sideband at ω_S, while the received signal is accompanied by an interferer at ω_{int}. Upon downconversion mixing, the desired channel is convolved with the carrier and the interferer with the sideband. If $\omega_{int} - \omega_S = \omega_0 - \omega_{LO}$ $(=\omega_{IF})$, then the downconverted interferer falls atop the desired channel. For example, if the interferer is 60 dB above the desired signal and the sideband is 70 dB below the carrier, then the corruption at the IF is 10 dB below the signal—barely an acceptable value in some standards.

Example 10.1

A receiver with an IIP_3 of -15 dBm senses a desired signal and two interferers as shown in Fig. 10.5. The LO also exhibits a sideband at ω_S, corrupting the downconversion. What relative LO sideband magnitude creates as much corruption as intermodulation does?

(Continues)

Example 10.1 (*Continued*)

Solution:

To compute the level of the resulting intermodulation product that falls into the desired channel, we write the difference between the interferer level and the IM_3 level in dB as

$$\Delta P = 2(IIP_3 - P_{in}) \tag{10.1}$$
$$= 50 \text{ dB}. \tag{10.2}$$

(The IM_3 level is equal to -90 dBm.) Thus, if the sideband is 50 dB below the carrier, then the two mechanisms lead to equal corruptions.

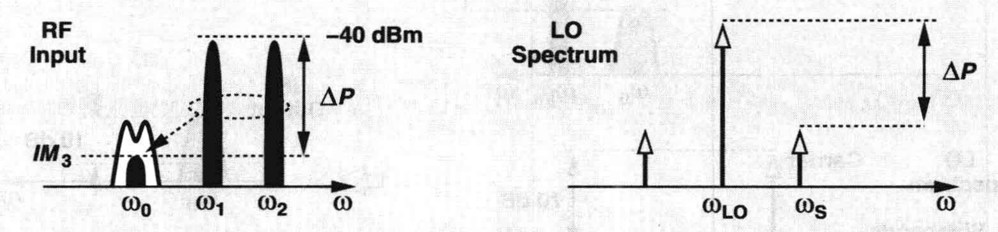


Figure 10.5 *Intermodulation and reciprocal mixing in a receiver.*

When the digital channel selection command in Fig. 10.3 changes value, the synthesizer takes a finite time to settle to a new output frequency (Fig. 10.6). Called the "lock time" for synthesizers that employ PLLs, this settling time directly subtracts from the time available for communication. The following example elaborates on this point.

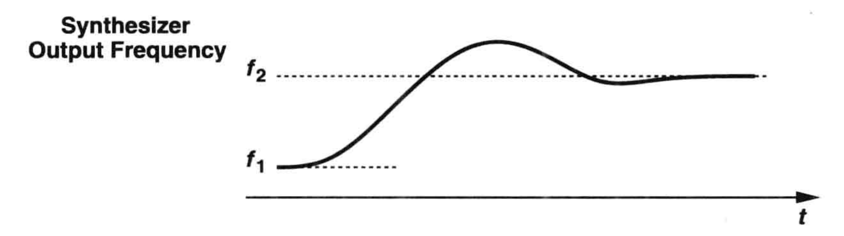


Figure 10.6 *Frequency settling during synthesizer lock period.*

Example 10.2

During synthesizer settling, the power amplifier in a transmitter is turned off. Explain why.

Solution:

If the power amplifier remains on, then the LO frequency variations produce large fluctuations in the transmitted carrier during the settling time. Shown in Fig. 10.7, this effect can considerably corrupt other users' channels.

Example 10.2 *(Continued)*

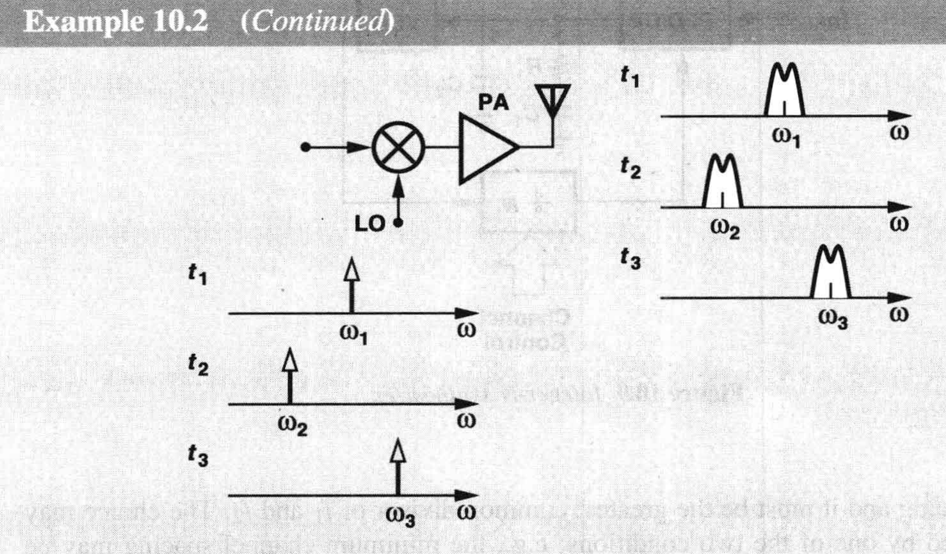


Figure 10.7 *Fluctuation of carrier frequency during synthesizer switching.*

Lock times required in typical RF systems vary from a few tens of milliseconds to a few tens of microseconds. (Exceptional cases such as ultra-wideband systems stipulate a lock time of less than 10 ns.) But how is the lock time defined? Illustrated in Fig. 10.8, the lock time is typically specified as the time required for the output frequency to reach within a certain margin (e.g., 100 ppm) around its final value.

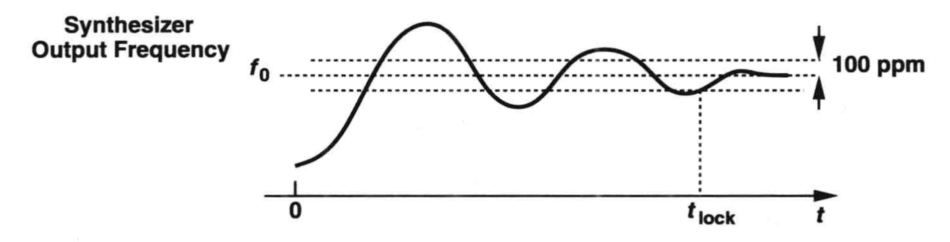


Figure 10.8 *Definition of synthesizer lock time.*

10.2 BASIC INTEGER-*N* SYNTHESIZER

Recall from Chapter 9 that a PLL employing a feedback divide ratio of N multiplies the input frequency by the same factor. Based on this concept, integer-N synthesizers produce an output frequency that is an *integer* multiple of the reference frequency (Fig. 10.9). If N increases by 1, then f_{out} increases by f_{REF}; i.e., the minimum channel spacing is equal to the reference frequency.

How does the synthesizer of Fig. 10.9 cover a desired frequency range, f_1 to f_2? The divide ratio must be programmable from, say, N_1 to N_2 so that $N_1 f_{REF} = f_1$ and $N_2 f_{REF} = f_2$. We therefore recognize two conditions for the choice of f_{REF}: it must be equal to the desired

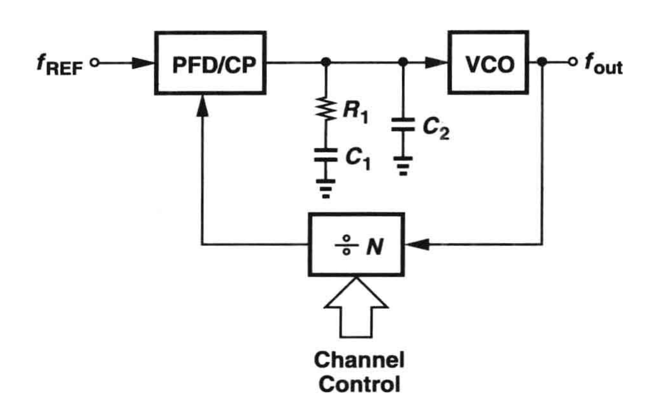


Figure 10.9 *Integer-N synthesizer.*

channel spacing and it must be the greatest common divisor of f_1 and f_2. The choice may be dominated by one of the two conditions; e.g., the minimum channel spacing may be smaller than the greatest common divisor of f_1 and f_2.

Example 10.3

Compute the required reference frequency and range of divide ratios for an integer-N synthesizer designed for a Bluetooth receiver. Consider two cases: (a) direct conversion, (b) sliding-IF downconversion with $f_{LO} = (2/3)f_{RF}$ (Chapter 4).

Solution:

(a) Shown in Fig. 10.10(a), the LO range extends from the *center* of the first channel, 2400.5 MHz, to that of the last, 2479.5 MHz. Thus, even though the channel spacing is 1 MHz, f_{REF} must be chosen equal to 500 kHz. Consequently, $N_1 = 4801$ and $N_2 = 4959$.

(b) As illustrated in Fig. 10.10(b), in this case the channel spacing and the center frequencies are multiplied by 2/3. Thus, $f_{REF} = 1/3$ MHz, $N_1 = 4801$, and $N_2 = 4959$.

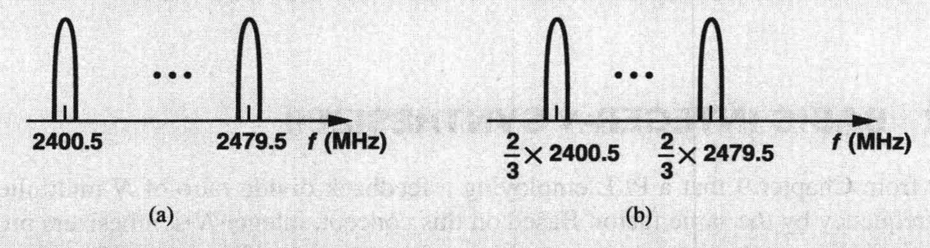


Figure 10.10 *Bluetooth LO frequency range for (a) direct and (b) sliding-IF downconversion.*

The simplicity of the integer-N synthesizer makes it an attractive choice. Behaving as a standard PLL, this architecture lends itself to the analyses carried out in Chapter 9.

In particular, the PFD/CP nonidealities and design techniques described in Chapter 9 directly apply to integer-N synthesizers.

10.3 SETTLING BEHAVIOR

Our study of PLL dynamics in Chapter 9 has dealt with frequency or phase changes at the *input*, a rare event in RF synthesizers. Instead, the transients of interest are those due to (1) a change in the feedback divide ratio, i.e., as the synthesizer hops from one channel to another, or (2) the startup switching, i.e., the synthesizer has been off to save power and is now turned on.

Let us consider the case of channel switching. Interestingly, a small change in N yields the same transient behavior as does a small change in the input frequency. This can be proved with the aid of the feedback system shown in Fig. 10.11, where the feedback factor A changes by a small amount ϵ at $t = 0$. The output after $t = 0$ is equal to

$$Y(s) = \frac{H(s)}{1 + (A + \epsilon)H(s)} X(s) \tag{10.3}$$

$$\approx \frac{H(s)}{1 + AH(s)} \cdot \frac{1}{1 + \epsilon/A} X(s) \tag{10.4}$$

$$\approx \frac{H(s)}{1 + AH(s)} \left(1 - \frac{\epsilon}{A}\right) X(s), \tag{10.5}$$

implying that the change is equivalent to multiplying $X(s)$ by $(1 - \epsilon/A)$ while retaining the same transfer function. Since in the synthesizer environment, $x(t)$ (the input frequency) is constant before $t = 0$, i.e., $x(t) = f_0$, we can view multiplication by $(1 - \epsilon/A)$ as a step function from f_0 to $f_0(1 - \epsilon/A)$,[1] i.e., a frequency jump of $-(\epsilon/A)f_0$.

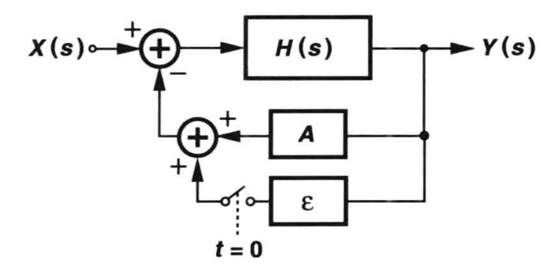


Figure 10.11 *Effect of changing the feedback factor.*

The foregoing analysis suggests that, when the divide ratio changes, the loop responds as if an input frequency step were applied, requiring a finite time to settle within an acceptable margin around its final value. As shown in Fig. 10.12, the worst case occurs when the synthesizer output frequency must go from the first channel, $N_1 f_{REF}$, to the last, $N_2 f_{REF}$, or vice versa.

1. Note that Eqs. (10.3)–(10.5) are written for frequency quantities, but they apply to phase quantities as well.

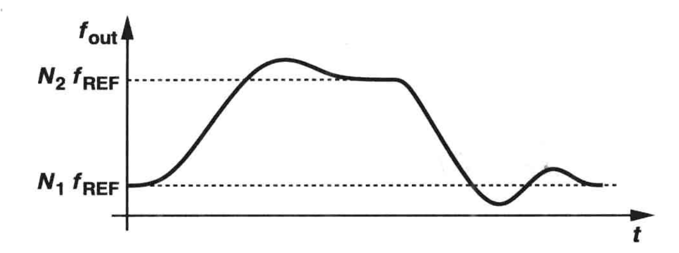


Figure 10.12 *Worst-case synthesizer settling.*

Example 10.4

In synthesizer settling, the quantity of interest is the frequency error, $\Delta\omega_{out}$, with respect to the final value. Determine the transfer function from the input frequency to this error.

Solution:

The error is equal to $\omega_{in}[N - H(s)]$, where $H(s)$ is the transfer function of a type-II PLL (Chapter 9). Thus,

$$\frac{\Delta\omega_{out}}{\omega_{in}} = N\frac{s^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}. \tag{10.6}$$

In order to estimate the settling time, we combine the above result with the equations derived in Chapter 9, assuming $N_2 - N_1 \ll N_1$. If the divide ratio jumps from N_1 to N_2, this change is equivalent to an input frequency step of $\Delta\omega_{in} = (N_2 - N_1)\omega_{REF}/N_1$.

We must also note that (a) the PLL settling equations are multiplied by the divide ratio, N_1 ($\approx N_2$) in an integer-N synthesizer, and (b) to obtain the settling time, the settling equations must be normalized to the final frequency, $N_2\omega_{REF}$. For the normalized error to fall below a certain amount, α, we have

$$\left|1 - \frac{N_1}{N_2}\right| g(t)u(t) \le \alpha, \tag{10.7}$$

where

$$g(t) = 1 - \left[\cos\left(\sqrt{1-\zeta^2}\omega_n t\right) - \frac{\zeta}{\sqrt{1-\zeta^2}}\sin\left(\sqrt{1-\zeta^2}\omega_n t\right)\right]e^{-\zeta\omega_n t} \quad \zeta < 1 \tag{10.8}$$

$$= 1 - (1 - \omega_n t)e^{-\zeta\omega_n t} \quad \zeta = 1 \tag{10.9}$$

$$= 1 - \left[\cosh\left(\sqrt{\zeta^2 - 1}\omega_n t\right) - \frac{\zeta}{\sqrt{\zeta^2-1}}\sinh\left(\sqrt{\zeta^2-1}\omega_n t\right)\right]e^{-\zeta\omega_n t} \quad \zeta > 1. \tag{10.10}$$

For example, if $\zeta = \sqrt{2}/2$, Eqs. (10.7) and (10.8) yield

$$\left|1 - \frac{N_1}{N_2}\right|\left(\cos\frac{\omega_n t_s}{\sqrt{2}} - \sin\frac{\omega_n t_s}{\sqrt{2}}\right)e^{-\omega_n t_s/\sqrt{2}} = \alpha, \tag{10.11}$$

where t_s denotes the settling time. A sufficient condition for the settling is that the exponential envelope decay to small values:

$$\left|1 - \frac{N_1}{N_2}\right| \sqrt{2} e^{-\omega_n t_s/\sqrt{2}} = \alpha, \tag{10.12}$$

where the factor of $\sqrt{2}$ represents the resultant of the cosine and the sine in Eq. (10.11). We therefore obtain the settling time for a normalized error of α as

$$t_s = \frac{\sqrt{2}}{\omega_n} \ln \left| \sqrt{2} \left(1 - \frac{N_1}{N_2}\right) \frac{1}{\alpha} \right|. \tag{10.13}$$

Example 10.5

A 900-MHz GSM synthesizer operates with $f_{REF} = 200$ kHz and provides 128 channels. If $\zeta = \sqrt{2}/2$, determine the settling time required for a frequency error of 10 ppm.

Solution:

The divide ratio is approximately equal to 4500 and varies by 128, i.e., $N_1 \approx 4500$ and $N_2 - N_1 = 128$. Equation (10.13) thus gives

$$t_s \approx \sqrt{2} \frac{8.3}{\omega_n}, \tag{10.14}$$

or

$$t_s = \frac{8.3}{\zeta \omega_n}. \tag{10.15}$$

While this relation has been derived for $\zeta = \sqrt{2}/2$, it provides a reasonable approximation for other values of ζ up to about unity.

How is the value of $\zeta \omega_n$ chosen? From Chapter 9, we note that the loop time constant is roughly equal to one-tenth of the input period. It follows that $(\zeta \omega_n)^{-1} \approx 10 T_{REF}$ and hence

$$t_s \approx 83 T_{REF}. \tag{10.16}$$

In practice, the settling time is longer and a rule of thumb for the settling of PLLs is 100 times the reference period.

As observed in Chapter 9 and in this section, the loop bandwidth trades with a number of critical parameters, including the settling time and the suppression of the VCO phase noise. Another important trade-off is that between the loop bandwidth and the magnitude of the reference sidebands. In the locked condition, the charge pump inevitably disturbs the control voltage at every phase comparison instant, modulating the VCO. In order to reduce this disturbance, the second capacitor (C_2 in Fig. 10.9) must be increased, and so must the main capacitor, C_1, because $C_2 \leq 0.2 C_1$.

The principal drawback of the integer-N architecture is that the output channel spacing is equal to the input reference frequency. We recognize that both the lock time (≈ 100 input cycles) and the loop bandwidth ($\approx 1/10$ of the input frequency) are tightly related to the channel spacing. Consequently, synthesizers designed for narrow-channel applications suffer from a long lock time and only slightly reduce the VCO phase noise.

10.4 SPUR REDUCTION TECHNIQUES

The trade-off between the loop bandwidth and the level of reference spurs has motivated extensive work on methods of spur reduction without sacrificing the bandwidth. Indeed, the techniques described in Chapter 9 alleviating issues such as charge sharing, channel-length modulation, and Up and Down current mismatch fall in this category. In this section, we study additional approaches that lower the ripple on the control voltage.

Example 10.6

A student reasons that if the transistor widths and drain currents in a charge pump are scaled down, so is the ripple. Is that true?

Solution:

This is true because the ripple is proportional to the *absolute* value of the unwanted charge pump injections rather than their relative value. This reasoning, however, can lead to the *wrong* conclusion that scaling the CP down reduces the output sideband level. Since a reduction in I_P must be compensated by a proportional increase in K_{VCO} so as to maintain ζ constant, the sideband level is almost unchanged.

A key point in devising spur reduction techniques is that the disturbance of the control voltage occurs primarily at the phase comparison instant. In other words, V_{cont} is disturbed for a short duration and remains relatively constant for the rest of the input period. We therefore surmise that the output sidebands can be lowered if V_{cont} is isolated from the disturbance for that duration. For example, consider the arrangement shown in Fig. 10.13(a), where S_1 turns off just before phase comparison begins and turns on slightly after the disturbance is finished. As a result, C_2 senses only the stable value at X and holds this value when S_1 is off. (The charge injection and clock feedthrough of S_1 still slightly disturb V_{cont}.)

Unfortunately, the arrangement of Fig. 10.13(a) leads to an unstable PLL. To understand this point, we recognize that the isolation of V_{cont} from the disturbance also eliminates the role of R_1. Since we keep S_1 off until the disturbance is finished, the voltage sensed by C_2 when S_1 turns on is *independent* of the value of R_1. That is, the circuit's behavior does not change if $R_1 = 0$. (As explained in Chapter 9, to create a zero, R_1 *must* produce a slight jump on the control voltage each time a finite, random phase error is detected.)

Let us now swap the two sections of the loop filter as shown in Fig. 10.13(b), where S_1 is still switched according to the waveforms in Fig. 10.13(a). Can this topology yield a stable PLL? Yes, it can. Upon experiencing a jump due to a finite phase error, node X delivers this jump to V_{cont} after S_1 turns on. Thus, the role of R_1 is maintained. On the other hand, the short-duration disturbance due to PFD and CP nonidealities is "masked"

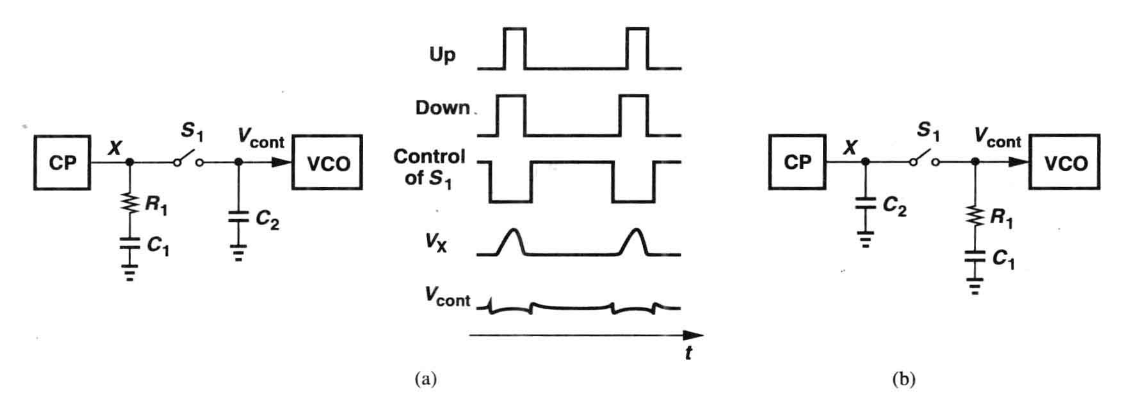


Figure 10.13 *Masking the ripple at node X by insertion of a switch, (a) with the second capacitor tied to V_{cont}, (b) with the main RC section tied to V_{cont}.*

by S_1, thereby leading to lower sidebands at the output [1, 2]. In practice, about half of C_2 is tied to V_{cont} so as to suppress the charge injection and clock feedthrough of S_1 [1, 2]. This "sampling loop filter" employs complementary transistors for S_1 to accommodate a rail-to-rail control voltage.

In order to arrive at another method of spur reduction, let us return to the open-loop transfer function of a type-II second-order PLL,

$$H_{open}(s) = \frac{I_P}{2\pi}\left(R_1 + \frac{1}{C_1 s}\right)\frac{K_{VCO}}{s} \tag{10.17}$$

and recall from Chapter 9 that R_1 is added in series with C_1 so as to create a zero in $H_{open}(s)$. We may then ask, is it possible to add a constant to K_{VCO}/s rather than to $1/(C_1 s)$? That is, can we realize

$$H_{open}(s) = \frac{I_P}{2\pi}\frac{1}{C_1 s}\left(\frac{K_{VCO}}{s} + K_1\right) \tag{10.18}$$

so as to obtain a zero? The loop now contains a zero at K_{VCO}/K_1, whose magnitude can be chosen to yield a reasonable damping factor.

Before computing the damping factor, we ponder the meaning of K_1 in Eq. (10.18). Since K_1 simply adds to the output *phase* of the VCO, we may surmise that it denotes a constant delay after the VCO. But the transfer function of such a delay [of the form $\exp(-K_1 s)$] would be *multiplied* by K_{VCO}/s. To avoid this confusion, we construct a block diagram representing Eq. (10.18) [Fig. 10.14(a)], recognizing that $K_{VCO}/s + K_1$ is, in fact, the transfer function from V_{cont} to ϕ_1, i.e.,

$$\frac{\phi_1}{V_{cont}}(s) = \frac{K_{VCO}}{s} + K_1. \tag{10.19}$$

That is, K_1 denotes a block that is *controlled* by V_{cont}. Indeed, K_1 represents a *variable-delay* stage [3] having a "gain" of K_1:

$$K_1 = \frac{\Delta T_d}{\Delta V_{cont}}, \tag{10.20}$$

where T_d is the delay of the stage [Fig. 10.14(b)].

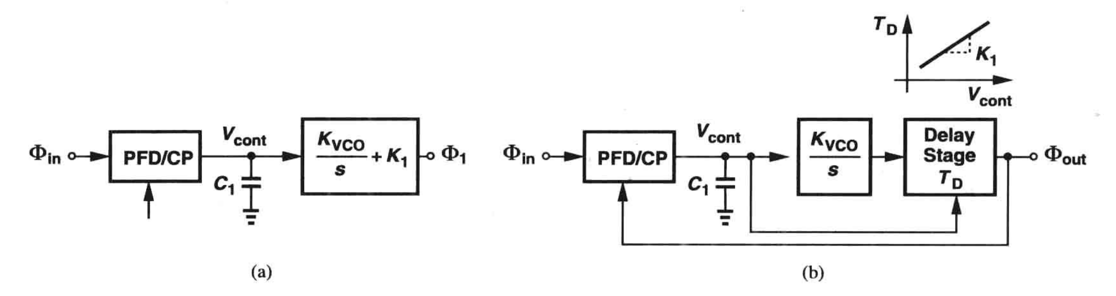


Figure 10.14 *(a) Stabilization of PLL by adding K_1 to the transfer function of VCO, (b) realization using a variable-delay stage.*

The key advantage of the topology shown in Fig. 10.14(b) over a standard type-II PLL is that, by avoiding the resistor in series with C_1, it allows this capacitor to absorb the PFD/CP nonidealities. By contrast, in the standard PLL, only the smaller capacitor plays this role.

In order to determine the damping factor, we use Eq. (10.18) to write the closed-loop transfer function:

$$H_{closed}(s) = \frac{\dfrac{I_P}{2\pi C_1 s}\left(\dfrac{K_{VCO}}{s} + K_1\right)}{1 + \dfrac{I_P}{2\pi C_1 s}\left(\dfrac{K_{VCO}}{s} + K_1\right)} \tag{10.21}$$

$$= \frac{\dfrac{I_P K_1}{2\pi C_1}s + \dfrac{I_P K_{VCO}}{2\pi C_1}}{s^2 + \dfrac{I_P K_1}{2\pi C_1}s + \dfrac{I_P K_{VCO}}{2\pi C_1}}. \tag{10.22}$$

It follows that

$$\zeta = \frac{K_1}{2}\sqrt{\frac{I_P}{2\pi C_1 K_{VCO}}} \tag{10.23}$$

$$\omega_n = \sqrt{\frac{I_P K_{VCO}}{2\pi C_1}}. \tag{10.24}$$

In Problem 10.1, we prove that, with a feedback divider, these parameters are revised as

$$\zeta = \frac{K_1}{2}\sqrt{\frac{I_P}{2\pi C_1 K_{VCO} N}} \tag{10.25}$$

$$\omega_n = \sqrt{\frac{I_P K_{VCO}}{2\pi C_1 N}}. \tag{10.26}$$

Equation (10.25) implies that K_1 must be scaled in proportion to N so as to maintain a reasonable value for ζ, a difficult task because delay stages that accommodate *high* frequencies inevitably exhibit a *short* delay. The architecture is therefore modified to that shown in Fig. 10.15, where the variable delay line appears *after* the divider [3]. The reader can show that

$$\zeta = \frac{K_1}{2}\sqrt{\frac{I_P N}{2\pi C_1 K_{VCO}}} \tag{10.27}$$

$$\omega_n = \sqrt{\frac{I_P K_{VCO}}{2\pi C_1 N}}. \tag{10.28}$$

A retiming flipflop can be inserted between the delay line and the PFD to remove the phase noise of the former (Section 10.6.7).

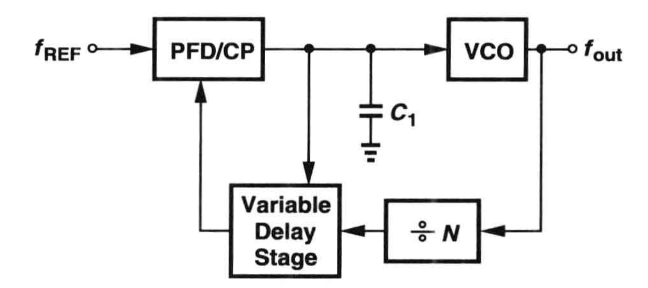


Figure 10.15 *Stabilization of an integer-N synthesizer.*

10.5 PLL-BASED MODULATION

In addition to the modulator and transmitter architectures introduced in Chapter 4, a number of other topologies can be realized that merge the modulation and frequency synthesis functions. We study two in this section and a few more in Chapter 12.

10.5.1 In-Loop Modulation

In addition to frequency synthesis, PLLs can also perform modulation. Recall from Chapter 3 that FSK and GMSK modulation can be realized by means of a VCO that senses the binary data. Figure 10.16(a) depicts a general case where the filter smoothes the time-domain transitions to some extent, thereby reducing the required bandwidth.[2] The principal issue here is the poor definition of the carrier frequency: the VCO center frequency drifts with time and temperature with no bound. One remedy is to phase-lock the VCO periodically to a reference so as to reset its center frequency. Illustrated in Fig. 10.16(b), such a system first disables the baseband data path and enables the PLL, allowing f_{out} to settle to Nf_{REF}. Next, the PLL is disabled and $x_{BB}(t)$ is applied to the VCO.

2. One may incorporate a simple analog filter even for FSK to improve the bandwidth efficiency.

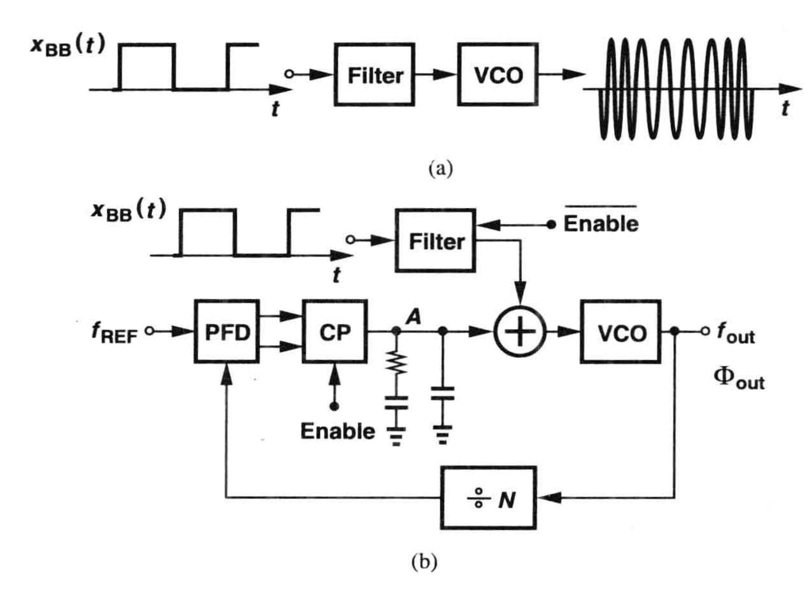


Figure 10.16 *(a) Open-loop modulation and (b) in-loop modulation of VCO.*

The arrangement of Fig. 10.16(b) requires periodic "idle" times during the communication to phase-lock the VCO, a serious drawback. Also, the output signal bandwidth depends on K_{VCO}, a poorly-controlled parameter. Moreover, the free-running VCO frequency may shift from Nf_{REF} due to a change in its load capacitance or supply voltage. Specifically, as depicted in Fig. 10.17, if the power amplifier is ramped at the beginning of transmission, its input impedance Z_{PA} changes considerably, thus altering the capacitance seen at the input of the buffer ("load pulling"). Also, upon turning on, the PA draws a very high current from the system supply, reducing its voltage by tens or perhaps hundreds of millivolts and changing the VCO frequency ("supply pushing").

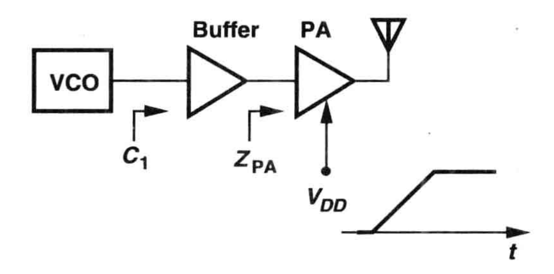


Figure 10.17 *Variation of buffer input impedance during PA ramp-up.*

To alleviate the foregoing issues, the VCO can remain *locked* while sensing the baseband data. That is, the PLL in Fig. 10.16(b) continuously monitors and corrects the VCO output (i.e., the CP is always enabled). Of course, to impress the data upon the carrier successfully, the design must select a very *slow* loop so that the desired phase modulation at the output is *not* corrected by the PLL. Called "in-loop modulation," this approach offers two advantages over the quadrature upconversion techniques studied in Chapter 4. First, in contrast to a quadrature GMSK modulator, it requires much less processing of the baseband data. Second, it obviates the need for the quadrature phases of the LO. Of course, this method can be applied only to constant-envelope modulation schemes.

Example 10.7

The effect of the PLL in Fig. 10.16(b) on the data can also be studied in the frequency domain. Neglecting the effect of the filter in the data path, determine the transfer function from $x_{BB}(t)$ to ϕ_{out}.

Solution:

Beginning from the output, we write the feedback signal arriving at the PFD as ϕ_{out}/N, subtract it from 0 (the input phase), and multiply the result by $I_P/(2\pi)[R_1 + (C_1 s)^{-1}]$, obtaining the signal at node A. We then add X_{BB} to this signal[3] and multiply the sum by K_{VCO}/s:

$$\left[-\frac{\phi_{out}}{N} \cdot \frac{I_P}{2\pi}\left(R_1 + \frac{1}{C_1 s}\right) + X_{BB}\right]\frac{K_{VCO}}{s} = \phi_{out}. \tag{10.29}$$

It follows that

$$\frac{\phi_{out}}{X_{BB}}(s) = \frac{K_{VCO}s}{s^2 + \frac{I_P K_{VCO} R_1}{2\pi N}s + \frac{I_P K_{VCO}}{2\pi N C_1}}. \tag{10.30}$$

This response is simply equal to the VCO phase noise transfer function (Chapter 9) multiplied by K_{VCO}/s. (Why is this result expected?) For low values of s, the system exhibits a high-pass response, attenuating the low-frequency contents of X_{BB}. As s becomes large enough that the denominator can be approximated by s^2, the response approaches the desired shape, K_{VCO}/s (that of a frequency modulator). Figure 10.18 plots this behavior. The reader can prove that the response reaches a peak equal to $K_{VCO}/(2\zeta\omega_n)$ at $\omega = \omega_n$. For the baseband data to experience negligible high-pass filtering, ω_n must be well below the lowest frequency content of the data. As a rule of thumb, we say ω_n should be around $1/1000$ of the bit rate.

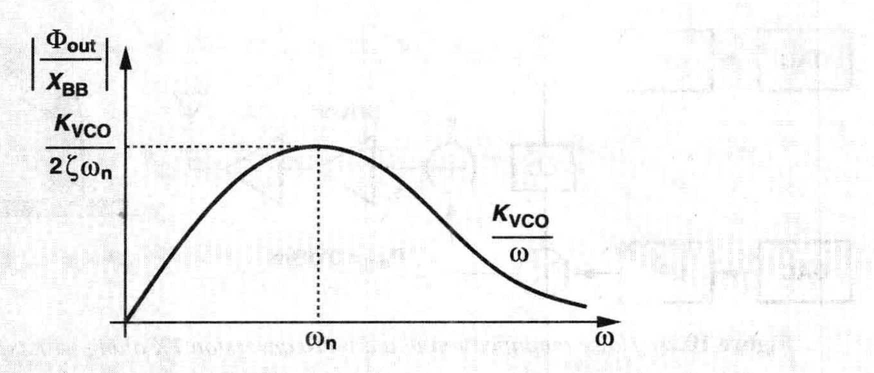


Figure 10.18 *In-loop modulation frequency response.*

In-loop modulation entails two drawbacks: (1) due to the very small PLL bandwidth, the VCO phase noise remains mostly uncorrected, and (2) the modulated signal bandwidth is a function of K_{VCO}, a process- and temperature-dependent parameter.

3. The effect of the filter in the X_{BB} path is neglected for simplicity.

10.5.2 Modulation by Offset PLLs

A stringent requirement imposed by GSM has led to a transmitter architecture that employs a PLL with "offset mixing." The requirement relates to the maximum noise that a GSM transmitter is allowed to emit in the GSM *receive* band, namely, −129 dBm/Hz. Illustrated in Fig. 10.19 is a situation where the TX noise becomes critical: user B receives a weak signal around f_1 from user A while user C, located in the close proximity of user B, transmits a high power around f_2 and significant broadband noise. As shown here, the noise transmitted by user C corrupts the desired signal around f_1.

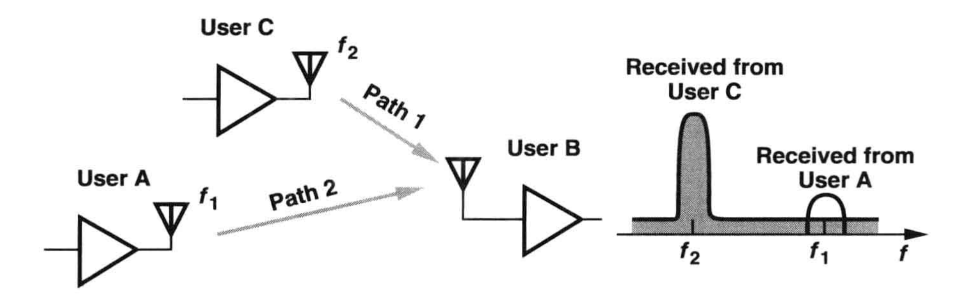


Figure 10.19 *Problem of transmitted noise in receiver band.*

The problem of broadband noise is particularly pronounced in direct-conversion transmitters. As depicted in Fig. 10.20, each stage in the signal path contributes noise, producing high output noise in the RX band even if the baseband LPF suppresses the out-of-channel DAC output noise. In Problem 10.2, we observe that the far-out phase noise of the LO also manifests itself as broadband noise at the PA output.

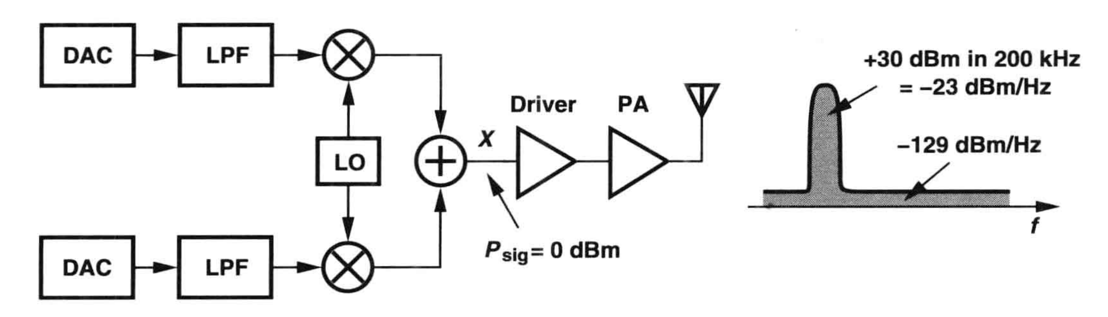


Figure 10.20 *Noise amplification in a direct-conversion TX along with typical values.*

Example 10.8

If the signal level is around 632 mV_{pp} (= 0 dBm in a 50-Ω system) at node X in Fig. 10.20, determine the maximum tolerable noise floor at this point. Assume the following stages are noiseless.

Example 10.8 *(Continued)*

Solution:

The noise floor must be 30 dB lower than that at the PA output, i.e., −159 dBm/Hz in a 50-Ω system (= 2.51 $nV_{rms}/\sqrt{Hz}$). Such a low level dictates very small load resistors for the upconversion mixers. In other words, it is simply impractical to maintain a sufficiently low noise floor at each point along the TX chain.

In order to reduce the TX noise in the RX band, a duplexer filter can be interposed between the antenna and the transceiver. (Recall from Chapter 4 that a duplexer is otherwise unnecessary in GSM because the TX and RX do not operate simultaneously.) However, the duplexer loss (2–3 dB) lowers the transmitted power and raises the receiver noise figure.

Alternatively, the upconversion chain can be modified so as to produce a small amount of broadband noise. For example, consider the topology shown in Fig. 10.21(a), where the

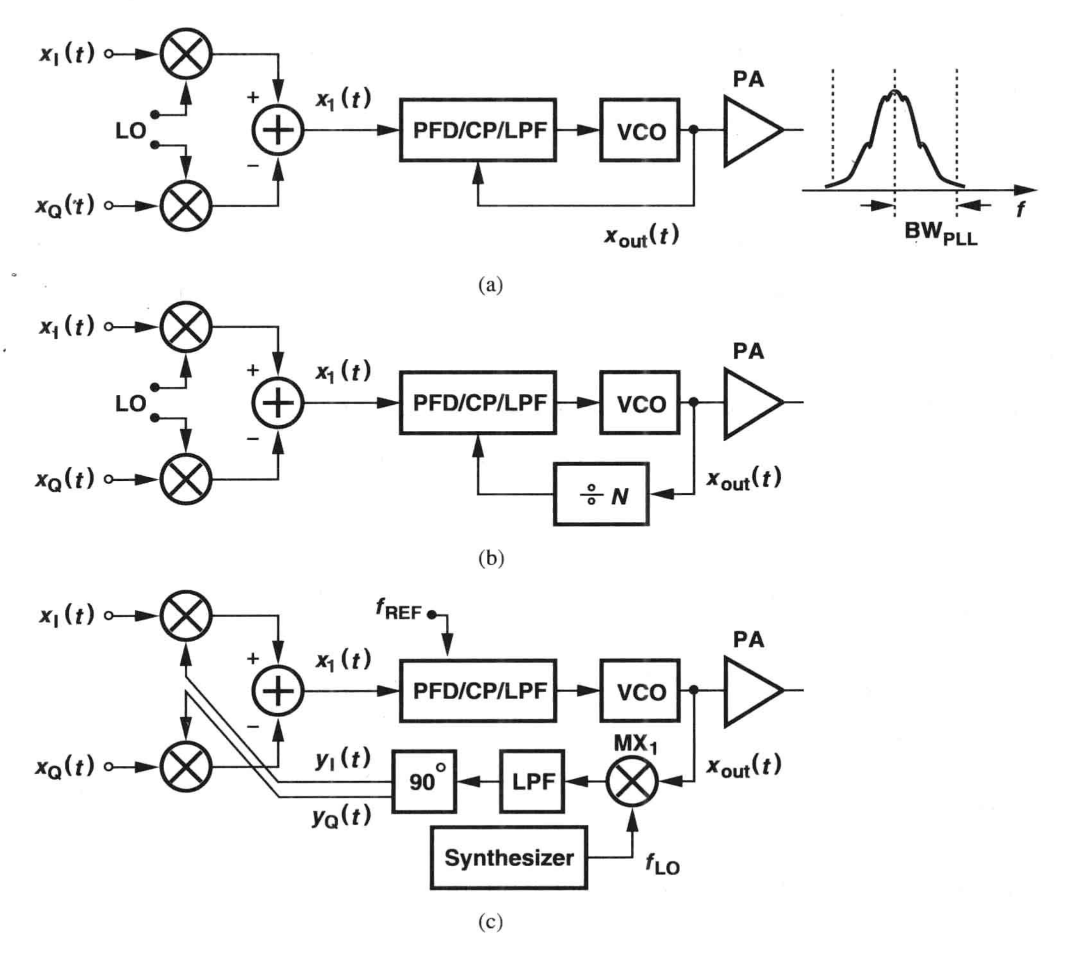


Figure 10.21 *(a) Noise filtration by means of a PLL, (b) use of ÷N in feedback, (c) offset-PLL architecture.*

baseband signal is upconverted and applied to a PLL. If the PLL bandwidth is only large enough to accommodate the signal, then $x_{out}(t) \approx x_1(t)$, but the broadband noise traveling to the antenna arises primarily from the far-out phase noise of the VCO. That is, unlike the TX chain in Fig. 10.20, this architecture need only minimize the broadband noise of one building block. Note that $x_1(t)$ has a constant envelope.

The above approach dictates that the PFD and CP operate at the carrier frequency, a relatively difficult requirement. We therefore add a feedback divider to the PLL to proportionally reduce the carrier frequency of $x_1(t)$ [Fig. 10.21(b)]. If $x_1(t) = A_1 \cos[\omega_1 t + \phi(t)]$, where $\phi(t)$ denotes GMSK or other types of frequency or phase modulation, and if the PLL bandwidth is large enough, then

$$x_{out}(t) = A_2 \cos[N\omega_1 t + N\phi(t)]. \tag{10.31}$$

Unfortunately, the PLL multiplies the phase by a factor of N, altering the signal bandwidth and modulation.

Let us now modify the architecture as shown in Fig. 10.21(c) [4]. Here, an "offset mixer," MX_1, downconverts the output to a center frequency of f_{REF}, and the result is separated into quadrature phases, mixed with the baseband signals, and applied to the PFD.[4] With the loop locked, $x_1(t)$ must become a faithful replica of the reference input, thus containing *no* modulation. Consequently, $y_I(t)$ and $y_Q(t)$ "absorb" the modulation information of the baseband signal. This architecture is called an "offset-PLL" transmitter or a "translational" loop.

Example 10.9

If $x_I(t) = A\cos[\phi(t)]$ and $x_Q(t) = A\sin[\phi(t)]$, derive expressions for $y_I(t)$ and $y_Q(t)$.

Solution:

Centered around f_{REF}, y_I and y_Q can be respectively expressed as

$$y_I(t) = a\cos[\omega_{REF}t + \phi_y(t)] \tag{10.32}$$

$$y_Q(t) = a\sin[\omega_{REF}t + \phi_y(t)], \tag{10.33}$$

where $\omega_{REF} = 2\pi f_{REF}$ and $\phi_y(t)$ denotes the phase modulation information. Carrying the quadrature upconversion operation and equating the result to an unmodulated tone, $x_1(t) = A\cos\omega_{REF}t$, we have

$$A_1 a\cos[\phi(t)]\cos[\omega_{REF}t + \phi_y(t)] - A_1 a\sin[\phi(t)]\sin[\omega_{REF}t + \phi_y(t)] = A\cos\omega_{REF}t. \tag{10.34}$$

It follows that

$$A_1 a\cos[\omega_{REF}t + \phi(t) + \phi_y(t)] = A\cos\omega_{REF}t \tag{10.35}$$

and hence

$$\phi_y(t) = -\phi(t). \tag{10.36}$$

Note that $x_{out}(t)$ also contains the same phase information.

4. The LPF removes the sum component at the output of MX_1.

The local oscillator waveform driving the offset mixer in Fig. 10.21(c) must, of course, be generated by another PLL according to the synthesis methods and concepts studied thus far in this chapter. However, the presence of two VCOs on the same chip raises concern with respect to mutual injection pulling between them. To ensure a sufficient difference between their frequencies, the offset frequency, f_{REF}, must be chosen high enough (e.g., 20% of f_{LO}). Additionally, two other reasons call for a large offset: (1) the stages following MX_1 must not degrade the phase margin of the overall loop, and (2) the center frequency of the 90° phase shift must be much greater than the signal bandwidth to allow accurate quadrature separation. Another variant of offset PLLs returns the output of the mixer directly to the PFD [5].

Example 10.10

In the architecture of Fig. 10.21(c), the PA output spectrum is centered around the VCO center frequency. Is the VCO injection-pulled by the PA?

Solution:

To the first order, it is not. This is because, unlike TX architectures studied in Chapter 4, this arrangement impresses the *same* modulated waveform on the VCO and the PA (Fig. 10.22). In other words, the instantaneous output voltage of the PA is simply an amplified replica of that of the VCO. Thus, the leakage from the PA arrives in-phase with the VCO waveform—as if a fraction of the VCO output were fed back to the VCO. In practice, the delay through the PA introduces some phase shift, but the overall effect on the VCO is typically negligible.

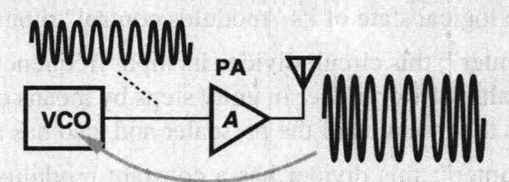


Figure 10.22 *Coupling of PA output to VCO in an offset-PLL TX.*

10.6 DIVIDER DESIGN

The feedback divider used in integer-N synthesizers presents interesting design challenges: (1) the divider modulus, N, must change in unity steps, (2) the first stage of the divider must operate as fast as the VCO, (3) the divider input capacitance and required input swing must be commensurate with the VCO drive capability, (4) the divider must consume low power, preferably less than the VCO. In this section, we describe divider designs that meet these requirements.

It is important to note that divider design typically assumes the VCO to have certain voltage swings and output drive capability and, as such, must be carried out in conjunction with the VCO design. Shown in Fig. 10.23 is an example where the VCO runs at twice the carrier frequency to avoid injection-pulling and is followed by a $\div 2$ stage. This divider may need to drive a considerable load capacitance, C_L, making it necessary to use wide

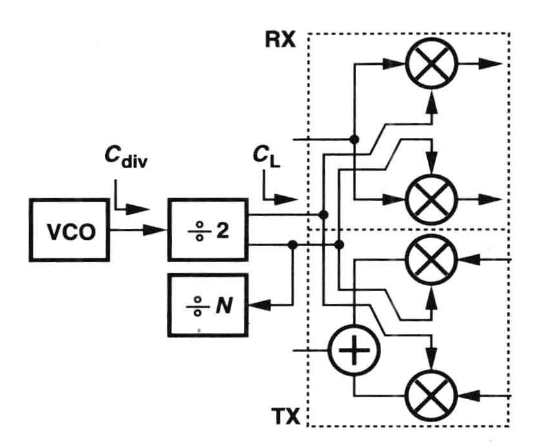


Figure 10.23 *Load seen by divider in a transceiver.*

transistors therein and hence present a large capacitance, C_{div}, to the VCO. A buffer can be inserted at the input and/or output of the divider but at the cost of greater power dissipation.

10.6.1 Pulse Swallow Divider

A common realization of the feedback divider that allows unity steps in the modulus is called the "pulse swallow divider." Shown in Fig. 10.24, the circuit consists of three blocks:

1. A "dual-modulus prescaler"; this counter provides a divide ratio of $N + 1$ or N according to the logical state of its "modulus control" input.
2. A "swallow counter"; this circuit divides its input frequency by a factor of S, which can be set to a value of 1 or higher in unity steps by means of the digital input.[5] This counter controls the modulus of the prescaler and also has a reset input.
3. A "program counter"; this divider has a constant modulus, P. When the program counter "fills up" (after it counts P pulses at its input), it resets the swallow counter.

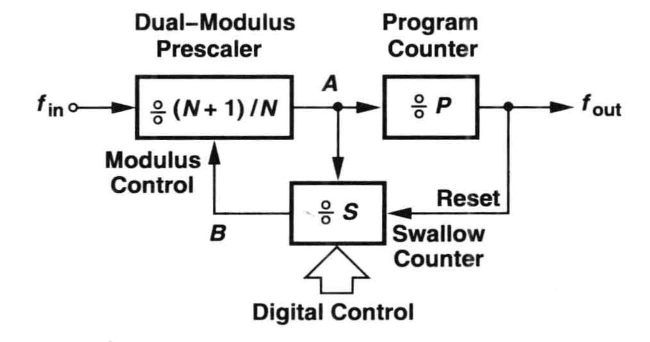


Figure 10.24 *Pulse swallow divider.*

5. It is unfortunate that the overall circuit is called the "pulse swallow divider" and this block, the "swallow counter."

We now show that the overall pulse swallow divider of Fig. 10.24(a) provides a divide ratio of $NP + S$. Suppose all three dividers begin from reset. The prescaler counts by $N + 1$, giving one pulse to the swallow counter (at point A) for every $N + 1$ pulses at the main input. The program counter counts the output pulses of the prescaler (point B). This continues until the swallow counter fills up, i.e., it receives S pulses at its input. [The main input therefore receives $(N + 1)S$ pulses.] The swallow counter then changes the modulus of the prescaler to N and begins from zero again. Note that the program counter has thus far counted S pulses, requiring another $P - S$ pulses to fill up. Now, the prescaler divides by N, producing $P - S$ pulses so as to fill up the program counter. In this mode, the main input must receive $N(P - S)$ pulses. Adding the total number of the pulses at the prescaler input in the two modes, we have $(N + 1)S + N(P - S) = NP + S$. That is, for every $NP + S$ pulses at the main input, the program counter generates one pulse at the output. The operation repeats after the swallow counter is reset. Note that P must be greater than S.

Sensing the high-frequency input, the prescaler proves the most challenging of the three building blocks. For this reason, numerous prescaler topologies have been introduced. We study some in the next section. As a rule of thumb, dual-modulus prescalers are about a factor of two slower than ÷2 circuits.

Example 10.11

In order to relax the speed required of the dual-modulus prescaler, the pulse swallow divider can be preceded by a ÷2 [Fig. 10.25(a)]. Explain the pros and cons of this approach.

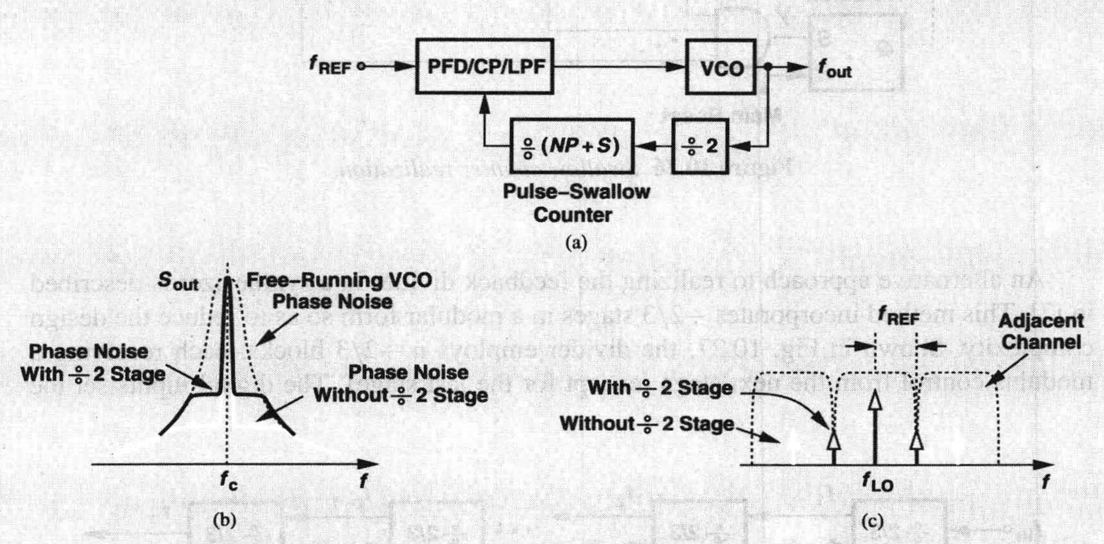


Figure 10.25 *(a) Use of ÷2 stage to relax the speed required of pulse swallow divider, (b) effect on output phase noise, (c) location of reference spurs with and without the ÷2 stage.*

Solution:

Here, $f_{out} = 2(NP + S)f_{REF}$. Thus, a channel spacing of f_{ch} dictates $f_{REF} = f_{ch}/2$. The lock speed and the loop bandwidth are therefore scaled down by a factor of two, making the

(Continues)

Example 10.11 *(Continued)*

VCO phase noise more pronounced [Fig. 10.25(b)]. One advantage of this approach is that the reference sideband lies at the edge of the adjacent channel rather than in the middle of it [Fig. 10.25(c)]. Mixed with little spurious energy, the sidebands can be quite larger than those in the standard architecture.

The swallow counter is typically designed as an asynchronous circuit for the sake of simplicity and power savings. Figure 10.26 shows a possible implementation, where cascaded ÷2 stages count the input and the NAND gates compare the count with the digital input, $D_n D_{n-1} \cdots D_1$. Once the count reaches the digital input, Y goes high, setting the RS latch. The latch output then disables the ÷2 stages. The circuit remains in this state until the main reset is asserted (by the program counter).

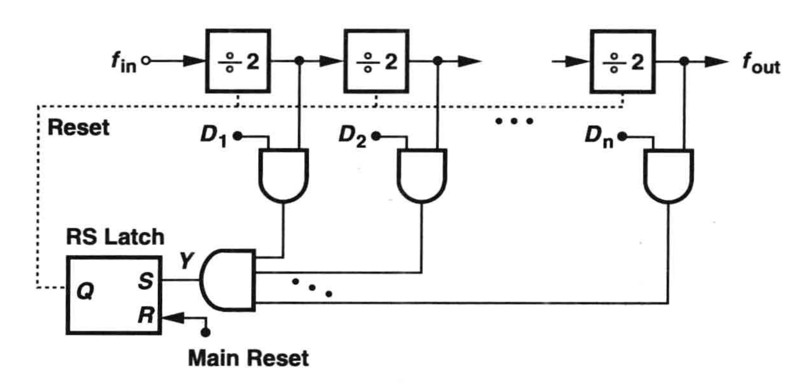


Figure 10.26 *Swallow counter realization.*

An alternative approach to realizing the feedback divider in a synthesizer is described in [7]. This method incorporates ÷2/3 stages in a modular form so as to reduce the design complexity. Shown in Fig. 10.27, the divider employs n ÷2/3 blocks, each receiving a modulus control from the next stage (except for the last stage). The digital inputs set the

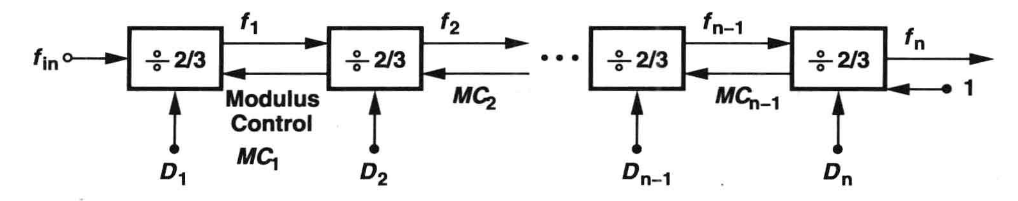


Figure 10.27 *Modular divider realizing multiple divide ratios.*

overall divide ratio according to

$$N = 2^n + D_n 2^{n-1} + D_{n-1} 2^{n-2} + \cdots + 2D_2 + D_1. \tag{10.37}$$

10.6.2 Dual-Modulus Dividers

As mentioned above, dual-modulus prescalers pose the most difficult challenge in divider design. We also note from our analysis of the pulse swallow divider in Section 10.6.1 that the modulus change must be *instantaneous*, an obvious condition but not necessarily met in all dual-modulus designs. As explained in the following sections, circuits such as the Miller divider and injection-locked dividers take a number of input cycles to reach the steady state.

Let us begin our study of dual-modulus prescalers with a divide-by-2/3 circuit. Recall from Chapter 4 that a ÷2 circuit can be realized as a D-flipflop placed in a negative feedback loop. A ÷3 circuit, on the other hand, requires two flipflops. Shown in Fig. 10.28 is an example,[6] where an AND gate applies $Q_1 \cdot \overline{Q_2}$ to the D input of FF_2. Suppose the circuit begins with $Q_1\overline{Q_2} = 00$. After the first clock, Q_1 assumes the value of $\overline{Q_2}$ (ZERO), and $\overline{Q_2}$ the value of $\overline{X}$ (ONE). In the next three cycles, $Q_1\overline{Q_2}$ goes to 10, 11, and 01. Note that the state $Q_1\overline{Q_2} = 00$ does not occur again because it would require the previous values of $\overline{Q_2}$ and X to be ZERO and ONE, respectively, a condition prohibited by the AND gate.

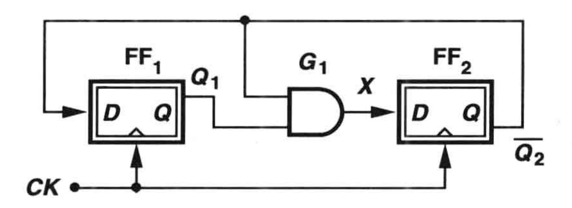


Figure 10.28 *Divide-by-3 circuit.*

Example 10.12

Design a ÷3 circuit using a NOR gate rather than an AND gate.

Solution:

We begin with the topology of Fig. 10.28, sense the Q output of FF_2, and add "bubbles" to compensate for the logical inversion [Fig. 10.29(a)]. The inversion at the input of FF_1 can now be moved to its output and hence realized as a bubble at the corresponding input of the AND gate [Fig. 10.29(b)]. Finally, the AND gate with two bubbles at its input can be replaced with a NOR gate [Fig. 10.29(c)]. The reader can prove that this circuit cycles through the following three states: $Q_1Q_2 = 00, 01, 10$.

(Continues)

6. In this book, we denote a latch by a single box and an FF by a double box.

Example 10.12 (*Continued*)

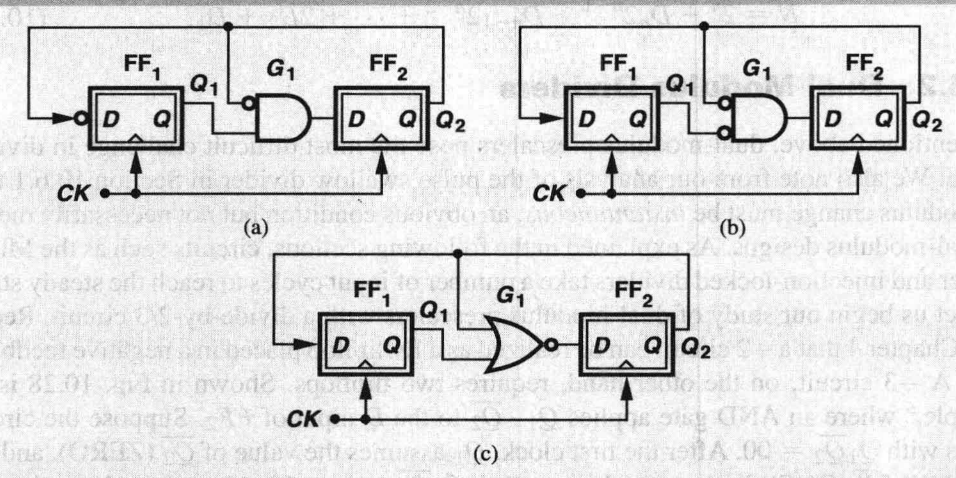


Figure 10.29 *Implementation of ÷3 circuit using a NOR gate: (a) use of Q_2 with a bubble at the NAND input and FF_1 input, (b) bubble moved from input of FF_1 to its output, (c) final realization.*

Example 10.13

Analyze the speed limitations of the ÷3 stage shown in Fig. 10.28.

Solution:

We draw the circuit as in Fig. 10.30(a), explicitly showing the two latches within FF_2. Suppose CK is initially low, L_1 is opaque (in the latch mode), and L_2 is transparent (in the sense mode). In other words, $\overline{Q_2}$ has just changed. When CK goes high and L_1 begins to sense, the value of $\overline{Q_2}$ must propagate through G_1 and L_1 before CK can fall again. Thus, the delay of G_1 enters the critical path. Moreover, L_2 must drive the input capacitance of FF_1, G_1, and an output buffer. These effects degrade the speed considerably, requiring that CK remain high long enough for $\overline{Q_2}$ to propagate to Y.

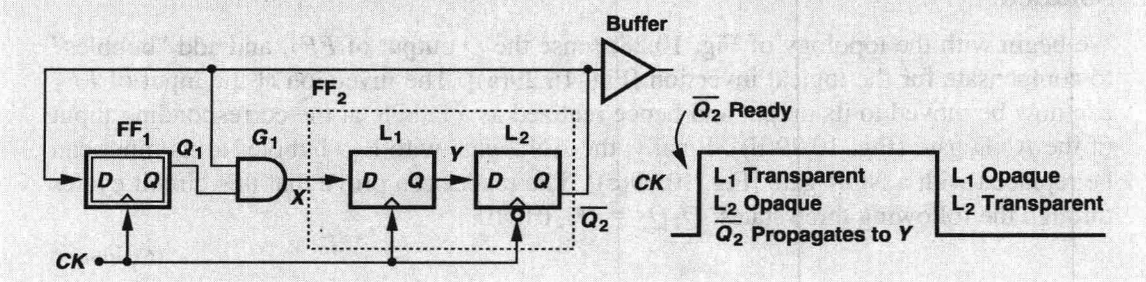


Figure 10.30 *Timing and critical path in ÷3 circuit.*

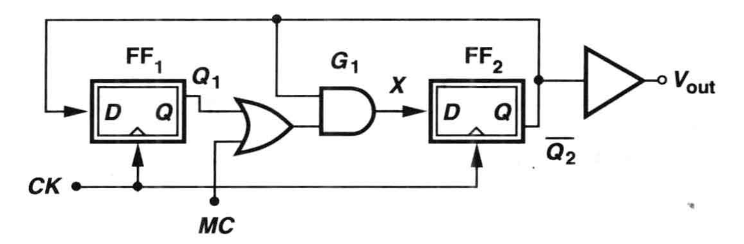


Figure 10.31 *Divide-by-2/3 circuit.*

The circuit of Fig. 10.28 can now be modified so as to have two moduli. Illustrated in Fig. 10.31, the ÷2/3 circuit employs an OR gate to permit ÷3 operation if the modulus control, MC, is low (why?) or ÷2 operation if it is high. In the latter case, only FF_2 divides the clock by 2 while FF_1 plays no role. Thus, the output can be provided by only FF_2.

Example 10.14

A student seeking a low-power prescaler design surmises that FF_1 in the circuit of Fig. 10.29 can be turned off when MC goes high. Explain whether this is a good idea.

Solution:

While saving power, turning off FF_1 may prohibit ***instantaneous*** modulus change because when FF_1 turns *on*, its initial state is undefined, possibly requiring an additional clock cycle to reach the desired value. For example, the overall circuit may begin with $Q_1\overline{Q_2} = 00$.

It is possible to rearrange the ÷2/3 stage so as to reduce the loading on the second flipflop. Illustrated in Fig. 10.32 [6], the circuit precedes each flipflop with a NOR gate. If MC is low, then $\overline{Q_1}$ is simply inverted by G_2—as if FF_2 directly followed FF_1. The circuit thus reduces to the ÷3 stage depicted in Fig. 10.29(c). If MC is high, Q_2 remains low, allowing G_1 and FF_1 to divide by two. Note that the output can be provided by only FF_1. This circuit has a 40% speed advantage over that in Fig. 10.31 [6].

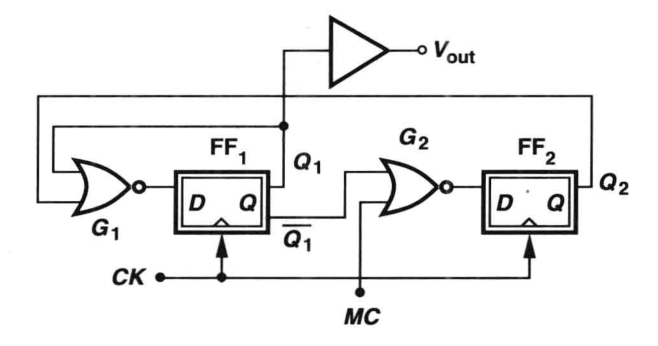


Figure 10.32 *Divide-by-2/3 circuit with higher speed.*

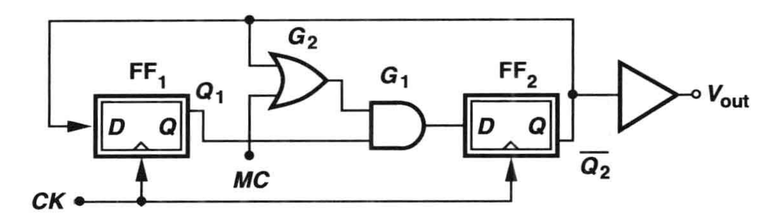


Figure 10.33 *Divide-by-3/4 circuit.*

Figure 10.33 shows a $\div 3/4$ stage. If MC = ONE, G_2 produces a ONE, allowing G_1 to simply pass the output of FF_1 to the D input of FF_2. The circuit thus resembles four latches in a loop and hence divides by 4. If $MC = 0$, G_2 passes $\overline{Q_2}$ to the input of G_1, reducing the circuit to that in Fig. 10.28. We observe that the critical path (around FF_2) contains a greater delay in this circuit than in the $\div 3$ stage of Fig. 10.28. A transistor-level design of such a divider is presented in Chapter 13.

The dual-modulus dividers studied thus far employ synchronous operation, i.e., the flipflops are clocked simultaneously. For higher moduli, a synchronous core having small moduli is combined with asynchronous divider stages. Figure 10.34 shows a $\div 8/9$ prescaler as an example. The $\div 2/3$ circuit (D23) of Fig. 10.31 is followed by two asynchronous $\div 2$ stages, and MC_1 is defined as the NAND of their outputs with the main modulus control, MC_2. If MC_2 is low, MC_1 is high, allowing D23 to divide by 2. The overall circuit thus operates as a $\div 8$ circuit. If MC_2 is high and the $\div 2$ stages begin from a reset state, then MC_1 is also high and D23 divides by 2. This continues until both A and B are high, at which point MC_1 falls, forcing D23 to divide by 3 for one clock cycle before A and B return to zero. The circuit therefore divides by 9 in this mode.

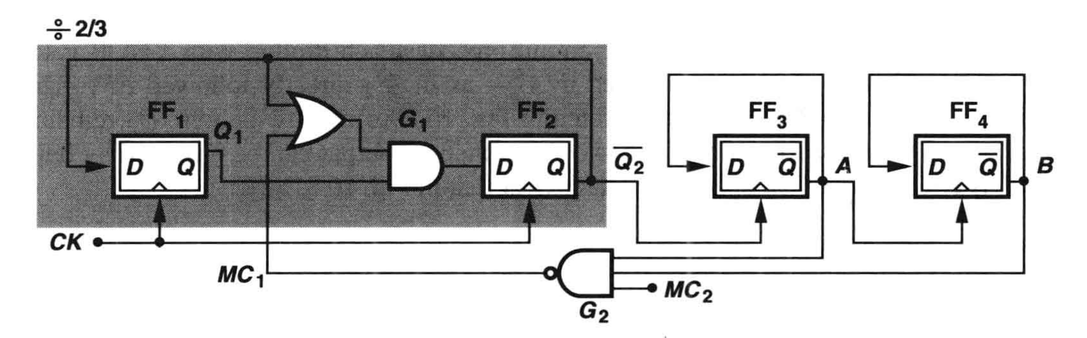


Figure 10.34 *Divide-by-8/9 circuit.*

Example 10.15

Design a $\div 15/16$ circuit using the synchronous $\div 3/4$ stage of Fig. 10.33.

Solution:

Since the $\div 3/4$ stage (D34) divides by 4 when MC is high, we surmise that only two more $\div 2$ circuits must follow to provide $\div 16$. To create $\div 15$, we must force D34 to divide by 3

Example 10.15 *(Continued)*

for one clock cycle. Shown in Fig. 10.35, the circuit senses the outputs of the asynchronous ÷2 stages by an OR gate and lowers MF when $AB = 00$. Thus, if MC is high, the circuit divides by 16. If MC is low and the ÷2 stages begin from 11, MF remains high and D34 divides by 4 until $AB = 00$. At this point, MF falls and D34 divides by 3 for one clock cycle before A goes high.

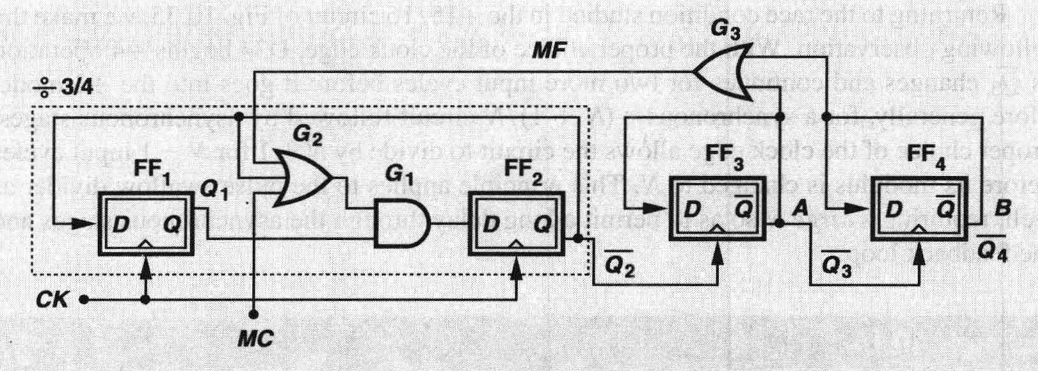
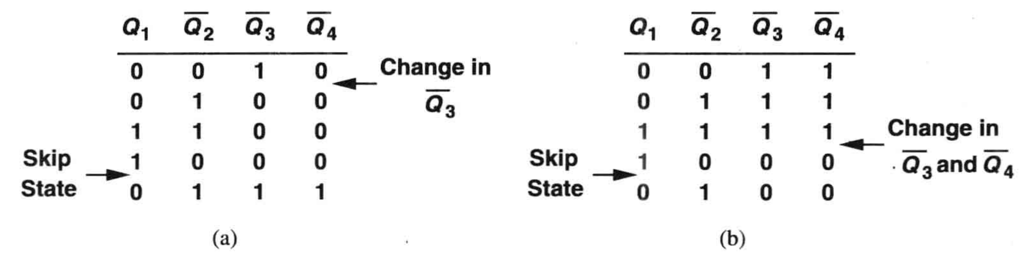


Figure 10.35 *Divide-by-15/16 circuit.*

An important issue in employing both synchronous and asynchronous sections in Fig. 10.35 is potential race conditions when the circuit divides by 15. To understand the problem, first suppose FF_3 and FF_4 change their output state on the *rising* edge of their clock inputs. If MC is low, the circuit continues to divide by 16, i.e., $Q_1\overline{Q_2}$ goes through the cycle: 01, 11, 10, 00 until both $\overline{Q_3}$ and $\overline{Q_4}$ are low. As depicted in Fig. 10.36(a), $Q_1\overline{Q_2}$ then skips the state 00 after the state 10. Since from the time $\overline{Q_3}$ goes low until the time $Q_1\overline{Q_2}$ skips one state, three CK_{in} cycles have passed, the propagation delay through FF_3 and G_3 need not be less than a cycle of CK_{in}.

Now consider a case where FF_3 and FF_4 change their output state on the *falling* edge of their clock inputs. Then, as shown in Fig. 10.36(b), immediately after $\overline{Q_3}\,\overline{Q_4}$ has fallen to 00, the ÷3/4 circuit must skip the state 00, mandating that the delay through FF_3, FF_4, and G_3 be less than half of a CK_{in} cycle. This is in general difficult to achieve, complicating the design and demanding higher power dissipation. Thus, the first choice is preferable.



(a)

Q_1	$\overline{Q_2}$	$\overline{Q_3}$	$\overline{Q_4}$
0	0	1	0
0	1	0	0
1	1	0	0
1	0	0	0
0	1	1	1

(b)

Q_1	$\overline{Q_2}$	$\overline{Q_3}$	$\overline{Q_4}$
0	0	1	1
0	1	1	1
1	1	1	1
1	0	0	0
0	1	0	0

Figure 10.36 *Delay budget in the ÷15/16 circuit with FF_3 and FF_4 activated on (a) rising edge, and (b) falling edge of clock.*

10.6.3 Choice of Prescaler Modulus

The pulse swallow divider of Fig. 10.24 provides a divide ratio of $NP + S$, allowing some flexibility in the choice of these three parameters. For example, to cover the Bluetooth channels from 2400 MHz to 2480 MHz, we can choose $N = 4$, $P = 575$, and $S = 100, \ldots, 180$, or $N = 10$, $P = 235$, and $S = 50, \ldots, 130$. (Recall that P must remain greater than S.) What trade-offs do we face in these choices? One aspect of the design calls for using a large N, and another for a small N.

Returning to the race condition studied in the $\div 15/16$ circuit of Fig. 10.35, we make the following observation. With the proper choice of the clock edge, D34 begins $\div 4$ operation as $\overline{Q_3}$ changes and continues for two more input cycles before it goes into the $\div 3$ mode. More generally, for a synchronous $\div(N+1)/N$ circuit followed by asynchronous stages, proper choice of the clock edge allows the circuit to divide by $N + 1$ for $N - 1$ input cycles before its modulus is changed to N. This principle applies to the pulse swallow divider as well, requiring a *large N* so as to permit a long delay through the asynchronous stages and the feedback loop.

Example 10.16

Consider the detailed view of a pulse swallow divider, shown in Fig. 10.37. Identify the critical feedback path through the swallow counter.

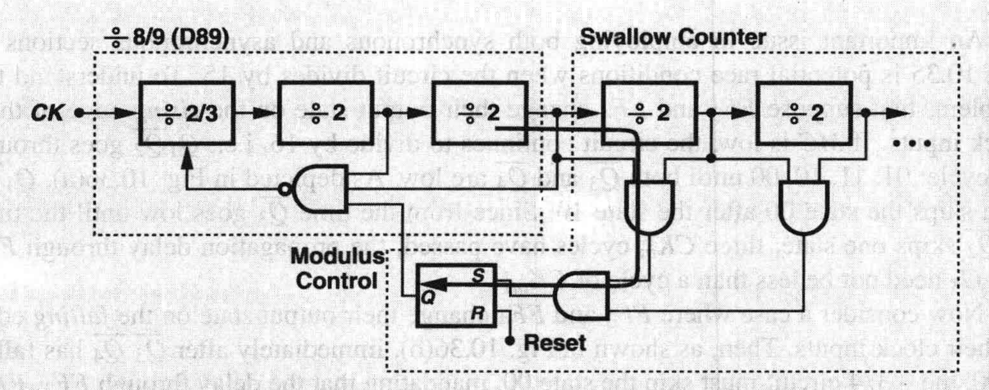


Figure 10.37 *Critical path in a pulse swallow divider.*

Solution:

When the $\div 9$ operation of the prescaler begins, the circuit has at most seven input cycles to change its modulus to 8. Thus, the last pulse generated by the prescaler in the previous $\div 8$ mode (just before the $\div 9$ mode begins) must propagate through the first $\div 2$ stage in the swallow counter, the subsequent logic, and the RS latch in fewer than seven input cycles.

The above perspective encourages a large N for the prescaler. On the other hand, a larger prescaler modulus leads to a higher power dissipation if the stages within the prescaler incorporate current steering to operate at high speeds (Section 10.6.4). For this reason, the prescaler modulus is determined by careful simulations. It is also possible to

pipeline the output of the RS latch in Fig. 10.37, thus allowing a smaller N but additional cycles for the modulus change [6].

10.6.4 Divider Logic Styles

The divider blocks in the feedback loop of a synthesizer can be realized by means of various logic styles. The choice of a divider topology is governed by several factors: the input swing (e.g., that available from the VCO), the input capacitance (e.g., that presented to the VCO), the maximum speed, the *output* swing (as required by the subsequent stage), the *minimum* speed (i.e., dynamic logic versus static logic), and the power dissipation. In this section, we study divider design in the context of different logic families.

Current-Steering Circuits Affording the fastest circuits, current-steering logic, also known as "current-mode logic" (CML), operates with moderate input and output swings. CML circuits provide differential outputs and hence a natural inversion; e.g., a single stage serves as both a NAND gate and an AND gate. CML derives its speed from the property that a differential pair can be rapidly enabled and disabled through its tail current source.

Figure 10.38(a) shows a CML AND/NAND gate. The top differential pair senses the differential inputs, A and $\overline{A}$, and is controlled by M_3 and hence by B and $\overline{B}$. If B is high, M_1 and M_2 remain on, $X = \overline{A}$, and $Y = A$. If B is low, M_1 and M_2 are off, X is at V_{DD}, and Y is pulled down by M_4 to $V_{DD} - R_D I_{SS}$. From another perspective, we note from Fig. 10.38(b) that M_1 and M_3 resemble a NAND branch, and M_2 and M_4 a NOR branch. The circuit is typically designed for a single-ended output swing of $R_D I_{SS} = 300$ mV, and the transistors are sized such that they experience complete switching with such input swings.

For the differential pairs in Fig. 10.38(a) to switch with moderate input swings, the transistors must not enter the triode region. For example, if M_3 is in the triode region when it is on, then the swings at B and $\overline{B}$ must be quite larger than 300 mV so as to turn M_3 off and M_4 on. Thus, the common-mode level of B and $\overline{B}$ must be below that of A and $\overline{A}$ by at least one overdrive voltage, making the design of the *preceding* stages difficult. Figure 10.39 depicts an example, where the NAND gate is preceded by two representative CML stages. Here, A and $\overline{A}$ swing between V_{DD} and $V_{DD} - R_1 I_{SS1}$. On the other hand, by virtue of the

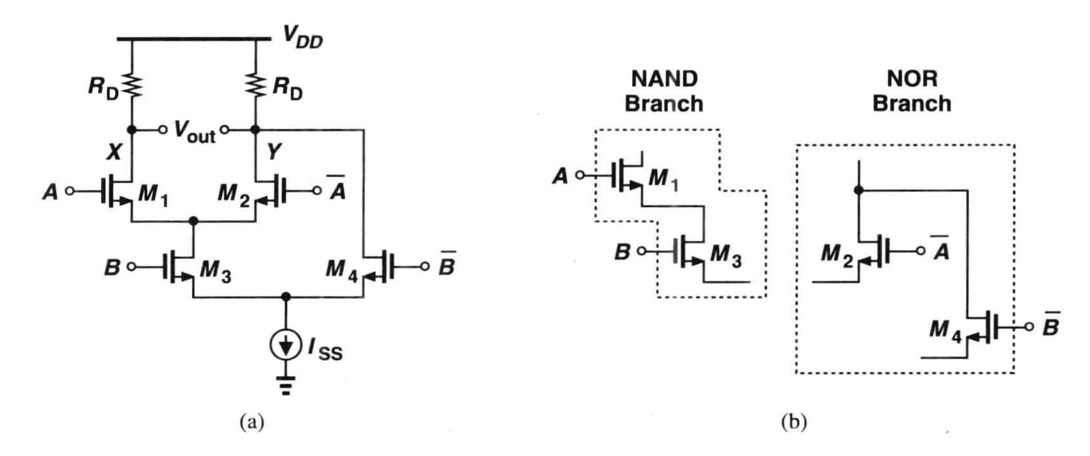


Figure 10.38 *(a) CML NAND realization, (b) NAND and NOR branches in the circuit.*

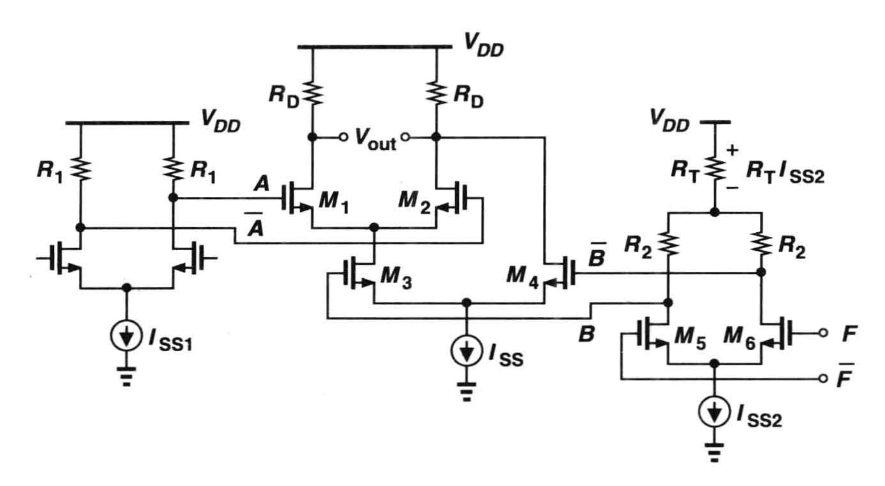


Figure 10.39 *Problem of common-mode compatibility at NAND inputs.*

level-shift resistor R_T, B and $\overline{B}$ vary between $V_{DD} - R_T I_{SS2}$ and $V_{DD} - R_T I_{SS2} - R_2 I_{SS2}$. That is, R_T shifts the CM level of B and $\overline{B}$ by $R_T I_{SS2}$. The addition of R_T appears simple, but now the *high* level of F and $\overline{F}$ is constrained if M_5 and M_6 must not enter the triode region. That is, this high level must not exceed $V_{DD} - R_T I_{SS2} - R_2 I_{SS2} + V_{TH}$.

The stacking of differential pairs in the NAND gate of Fig. 10.38(a) does not lend itself to low supply voltages. The CML NOR/OR gate, on the other hand, avoids stacking. Shown in Fig. 10.40(a), the circuit steers the tail current to the left if A or B is high, producing a low level at X and a high level at Y. Unfortunately, however, this stage operates only with single-ended inputs, demanding great attention to the CM level of A and B and the choice of V_b. Specifically, V_b must be generated such that it *tracks* the CM level of A and B. As illustrated in Fig. 10.40(b), V_b is established by a branch replicating the circuitry that produces A. The CM level of A is equal to $V_{DD} - R_2 I_{SS2}/2$, and so is the value of V_b. For very high speeds, a capacitor may be tied from V_b to V_{DD}, thereby maintaining a solid ac ground at the gate of M_3.

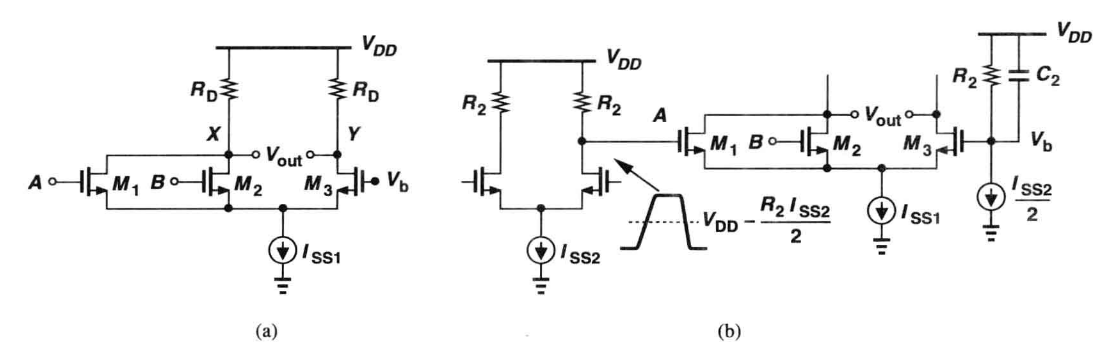


Figure 10.40 *(a) CML NOR gate, (b) proper generation of bias voltage V_b.*

At low supply voltages, we design the logic in the dividers to incorporate the NOR gate of Fig. 10.40(a) rather than the NAND circuit of Fig. 10.38(a). The ÷2/3 circuit of Fig. 10.32 exemplifies this principle. To ensure complete switching of M_1-M_3 in the NOR stage, the input swings must be somewhat larger than our rule of thumb of 300 mV, or the transistors must be wider.

Example 10.17

Should M_1-M_3 in Fig. 10.40(a) have equal widths?

Solution:

One may postulate that, if both M_1 and M_2 are on, they operate as a single transistor and absorb all of I_{SS1}, i.e., W_1 and W_2 need not exceed $W_3/2$. However, the worst case occurs if only M_1 or M_2 is on. Thus, for either transistor to "overcome" M_3, we require that $W_1 = W_2 \geq W_3$.

Another commonly-used gate is the XOR circuit, shown in Fig. 10.41. The topology is identical to the Gilbert cell mixer studied in Chapter 6, except that both input ports are driven by large swings to ensure complete switching. As with the CML NAND gate, this circuit requires proper CM level shift for B and $\overline{B}$ and does not easily operate with low supply voltages.

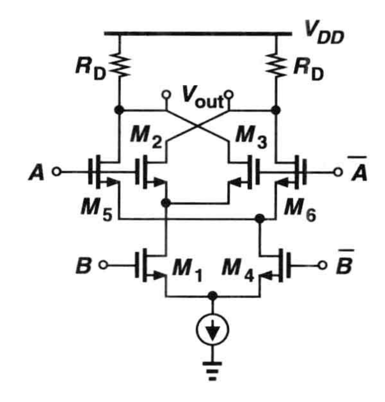


Figure 10.41 *CML XOR implementation.*

Figure 10.42 depicts an XOR gate that avoids stacking [8]. If A or B is high, M_3 turns off; that is, $I_{D3} = \overline{A + B}$. Similarly, $I_{D6} = \overline{\overline{A} + \overline{B}}$. The summation of I_{D3} and I_{D6} at node X is equivalent to an OR operation, and the flow of the sum through R_D produces an inversion.

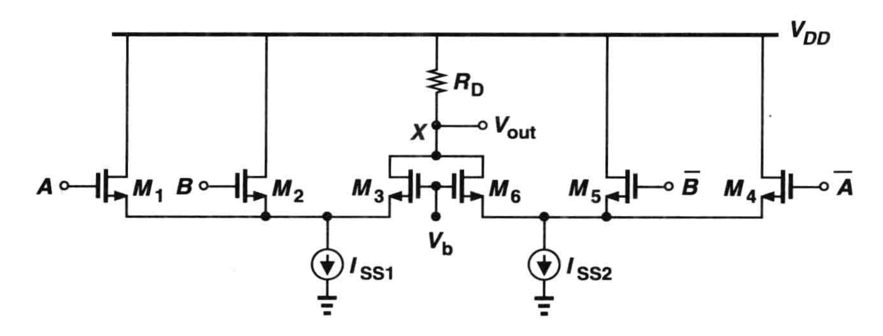


Figure 10.42 *Symmetric, low-voltage XOR.*

Thus,

$$V_{out} = \overline{(\overline{A + B} + \overline{\overline{A} + \overline{B}})} \tag{10.38}$$

$$= \overline{A}B + A\overline{B}. \tag{10.39}$$

In contrast to the XOR gate of Fig. 10.41, this circuit exhibits perfect symmetry with respect to A and B, an attribute that proves useful in some applications.

While lending itself to low supply voltages, the XOR topology of Fig. 10.42 senses each of the inputs in single-ended form, facing issues similar to those of the NOR gate of Fig. 10.40(a). In other words, V_b must be defined carefully and the input voltage swings and/or transistor widths must be larger than those required for the XOR of Fig. 10.41. Also, to provide differential outputs, the circuit must be duplicated with A and $\overline{A}$ (or B and $\overline{B}$) swapped.

The speed advantage of CML circuits is especially pronounced in latches. Figure 10.43(a) shows a CML D latch. The circuit consists of an input differential pair, M_1-M_2, a latch or "regenerative" pair, M_3-M_4, and a clocked pair, M_5-M_6. In the "sense mode," CK is high, and M_5 is on, allowing M_1-M_2 to sense and amplify the difference between D and $\overline{D}$. That is, X and Y track the input. In the transition to the "latch mode" (or "regeneration mode"), CK goes down, turning M_1-M_2 off, and $\overline{CK}$ goes up, turning M_3-M_4 on. The circuit now reduces to that in Fig. 10.43(b), where the positive feedback around M_3 and M_4 regeneratively amplifies the difference between V_X and V_Y. If the loop gain exceeds unity, the regeneration continues until one transistor turns off, e.g., V_X rises to V_{DD} and V_Y falls to $V_{DD} - R_D I_{SS}$. This state is retained until CK changes and the next sense mode begins.

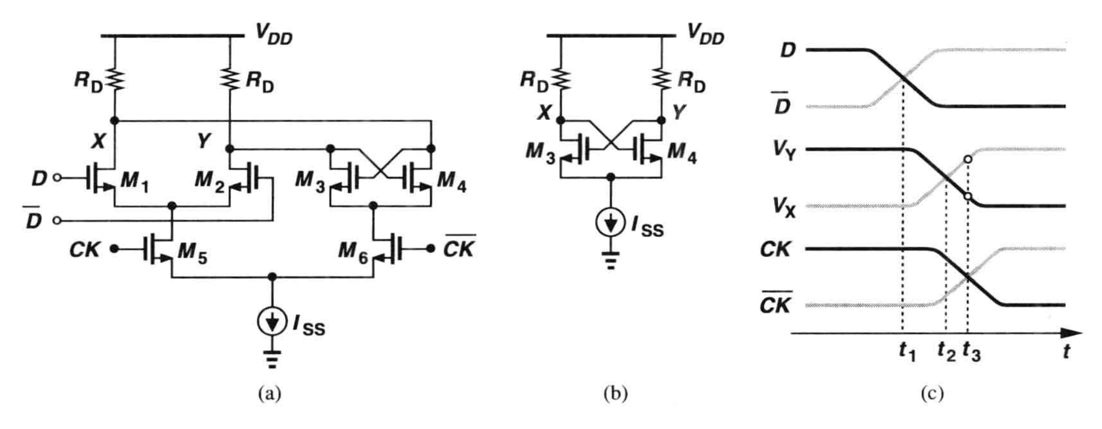


Figure 10.43 *(a) CML latch, (b) circuit in regeneration mode, (c) circuit's waveforms.*

In order to understand the speed attributes of the latch, let us examine its voltage waveforms as the circuit goes from the sense mode to the latch mode. As shown in Fig. 10.43(c), D and $\overline{D}$ cross at $t = t_1$, and V_X and V_Y at $t = t_2$. Even though V_X and V_Y have not reached their full swings at $t = t_3$, the circuit can enter the latch mode because the regenerative pair continues the amplification after $t = t_3$. Of course, the latch mode must be long enough for V_X and V_Y to approach their final values. We thus conclude that the latch operates properly

even with a limited bandwidth at X and Y if (a) in the sense mode, V_X and V_Y begin from their full levels and *cross*, and (b) in the latch mode, the initial difference between V_X and V_Y can be amplified to a final value of $I_{SS}R_D$.

Example 10.18

Formulate the regenerative amplification of the circuit in Fig. 10.43(b) if $V_X - V_Y$ begins with an initial value of V_{XY0}.

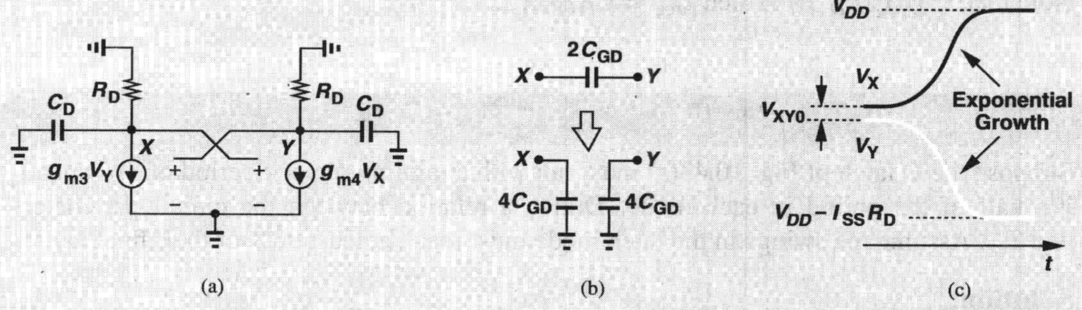


Figure 10.44 *(a) CML latch in regeneration mode, (b) decomposition of C_{GD}, and (c) circuit's waveforms.*

Solution:

If V_{XY0} is small, M_3 and M_4 are near equilibrium and the small-signal equivalent circuit can be constructed as shown in Fig. 10.44(a). Here, C_D represents the total capacitance seen at X and Y to ground, including $C_{GD1}+C_{DB1}+C_{GS3}+C_{DB3}+4C_{GD3}$ and the input capacitance of the next stage. The gate-drain capacitance is multiplied by a factor of 4 because it arises from both M_3 and M_4 and it is driven by differential voltages [Fig. 10.44(b)]. Writing a KCL at node X gives

$$\frac{V_X}{R_D} + C_D\frac{dV_X}{dt} + g_{m3,4}V_Y = 0. \tag{10.40}$$

Similarly,

$$\frac{V_Y}{R_D} + C_D\frac{dV_Y}{dt} + g_{m3,4}V_X = 0. \tag{10.41}$$

Subtracting (10.41) from (10.40) and grouping the terms, we have

$$-R_DC_D\frac{d(V_X - V_Y)}{dt} = (1 - g_{m3,4}R_D)(V_X - V_Y). \tag{10.42}$$

We denote $V_X - V_Y$ by V_{XY}, divide both sides of Eq. (10.42) by $-R_DC_DV_{XY}$, multiply both sides by dt, and integrate with the initial condition $V_{XY}(t = 0) = V_{XY0}$. Thus,

$$V_{XY} = V_{XY0}\exp\frac{(g_{m3,4}R_D - 1)t}{R_DC_D}. \tag{10.43}$$

(Continues)

Example 10.18 *(Continued)*

Interestingly, V_{XY} grows exponentially with time [Fig. 10.44(c)], exhibiting a "regeneration time constant" of

$$\tau_{reg} = \frac{R_D C_D}{g_{m3,4} R_D - 1}. \tag{10.44}$$

Of course, as V_{XY} increases, one transistor begins to turn off and its g_m falls toward zero. Note that, if $g_{m3,4} R_D \gg 1$, then $\tau_{reg} \approx C_D / g_{m3,4}$.

Example 10.19

Suppose the D latch of Fig. 10.43(a) must run with a minimum clock period of T_{ck}, spending half of the period in each mode. Derive a relation between the circuit parameters and T_{ck}. Assume the swings in the latch mode must reach at least 90% of their final value.

Solution:

We begin our calculation in the regeneration mode. Since the regenerative pair must produce $V_{XY} = 0.9 I_{SS} R_D$ in $0.5T_{ck}$ seconds, it requires an initial voltage difference, V_{XY0}, that can be obtained from (10.43):

$$V_{XY0} = 0.9 I_{SS} R_D \exp \frac{0.5 T_{ck}}{\tau_{reg}}. \tag{10.45}$$

The minimum initial voltage must be established by the input differential pair in the sense mode [just before $t = t_3$ in Fig. 10.43(c)]. In the worst case, when the sense mode begins, V_X and V_Y are at the opposite extremes and must cross and reach V_{XY0} in $0.5T_{ck}$ seconds (Fig. 10.45). For example, V_Y begins at V_{DD} and falls according to

$$V_Y(t) = V_{DD} - I_{SS} R_D \left(1 - \exp \frac{-t}{R_D C_D}\right). \tag{10.46}$$

Similarly,

$$V_X(t) = V_{DD} - I_{SS} R_D \exp \frac{-t}{R_D C_D}. \tag{10.47}$$

Since $V_X - V_Y$ must reach Eq. (10.45) in $0.5T_{ck}$ seconds, we have

$$-2 I_{SS} R_D \exp \frac{-0.5 T_{ck}}{R_D C_D} + I_{SS} R_D = 0.9 I_{SS} R_D \exp \frac{-0.5 T_{ck}}{\tau_{reg}}, \tag{10.48}$$

and hence

$$0.9 \exp \frac{-0.5 T_{ck}}{\tau_{reg}} + 2 \exp \frac{-0.5 T_{ck}}{R_D C_D} = 1. \tag{10.49}$$

With T_{ck} known, this expression constrains the upper bound on $R_D C_D$ and the lower bound on $g_{m3,4}$. In practice, the finite rise and fall times of the clock leave less than $0.5T_{ck}$ for each mode, tightening these constraints.

Example 10.19 *(Continued)*

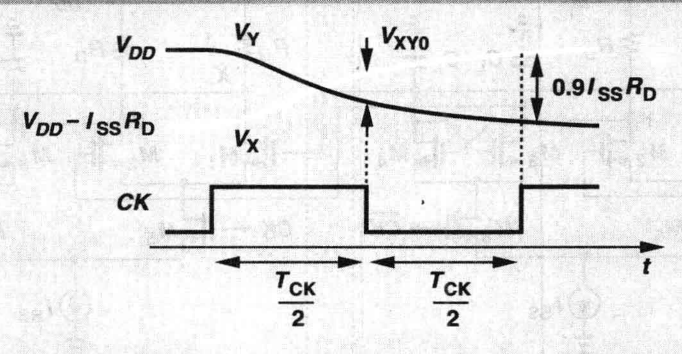


Figure 10.45 *Latch waveforms showing minimum time required for proper operation.*

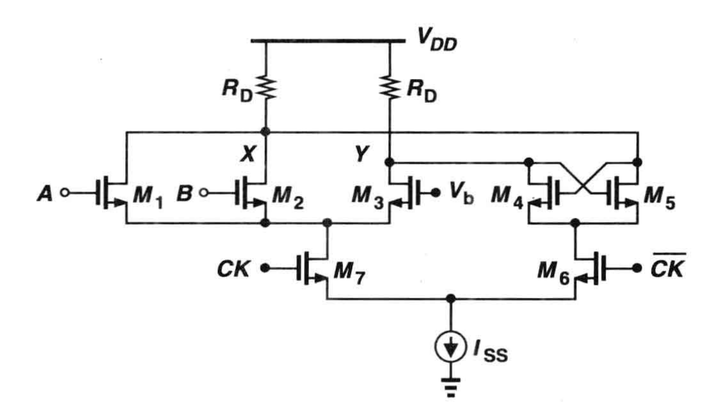


Figure 10.46 *CML latch incorporating NOR gate.*

It is possible to merge logic with a latch, thus reducing both the delay and the power dissipation. For example, the NOR and the master latch of FF_1 depicted in Fig. 10.32 can be realized as shown in Fig. 10.46. The circuit performs a NOR/OR operation on A and B in the sense mode and stores the result in the latch mode.

Design Procedure Let us construct a $\div 2$ circuit by placing two D latches in a negative feedback loop (Fig. 10.47). Note that the total capacitance seen at the clock input is twice that of a single latch. The design of the circuit begins with three known parameters: the power budget, the clock swing, and the load capacitance (the input capacitance of the next stage). We then follow these steps: (1) Select I_{SS} based on the power budget; (2) Select $R_D I_{SS} \approx 300$ mV; (3) Select $(W/L)_{1,2}$ such that the differential pair experiences nearly complete switching for a differential input of 300 mV; (4) Select $(W/L)_{3,4}$ such that the small-signal gain around the regenerative loop exceeds unity; (5) Select $(W/L)_{5,6}$ such that the clocked pair steers most of the tail current with the specified clock swing. It is important to ensure that the feedback around the loop is negative (why?).

The rough design thus obtained, along with the specified load capacitance, reaches a speed higher than the limit predicted by Eq. (10.49) because the voltage swings at X

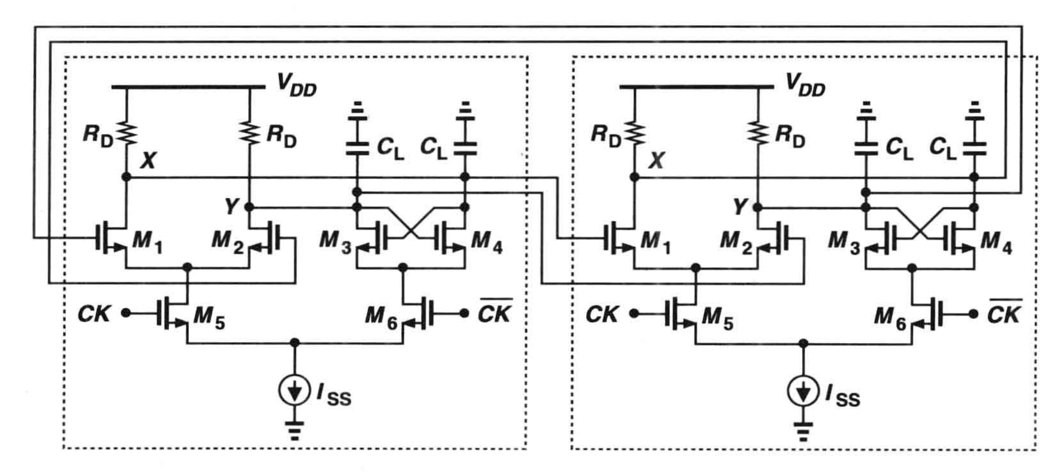


Figure 10.47 *Divide-by-2 circuit consisting of two CML latches in a negative feedback loop.*

and Y in each latch need not reach the full amount of $I_{SS}R_D$ for proper operation. In other words, as the clock frequency exceeds the limit given by (10.49), the output swings become smaller—up to a point where V_X and V_Y simply do not have enough time to cross and the circuit fails.

In practice, the transistor widths may need to exceed those obtained above for three reasons: (a) the tail node voltages of M_1-M_2 and M_3-M_4 may be excessively low, driving M_5-M_6 into the triode region; (b) the tail node voltage of M_5-M_6 may be so low as to leave little headroom for I_{SS}; and (c) at very high speeds, the voltage swings at X and Y do not reach R_DI_{SS}, demanding wider transistors for steering the currents.

Example 10.20

The performance of high-speed dividers is typically characterized by plotting the minimum required clock voltage swing ("sensitivity") as a function of the clock frequency. Sketch the sensitivity for the ÷2 circuit of Fig. 10.47.

Solution:

For a clock with abrupt edges, we expect the required clock swing to remain relatively constant up to the point where the internal time constants begin to manifest themselves. Beyond this point, the required swing must increase. The overall behavior, however, appears as shown in Fig. 10.48. Interestingly, the required clock swing falls to *zero* at some frequency, f_1. Since for zero input swings, I_{SS} is simply split equally between M_5 and M_6 in Fig. 10.47, the circuit reduces to that depicted in Fig. 10.49. We recognize that the result resembles a two-stage *ring oscillator*. In other words, in the absence of an input clock, the circuit simply oscillates at a frequency of $f_1/2$. This observation provides another perspective on the operation of the divider: the circuit behaves as an oscillator that is injection-locked to the input clock (Section 10.6.6). This viewpoint also explains why the clock swing cannot be arbitrarily small at low frequencies. Even with square clock waveforms, a small swing fails to steer all of the tail current, thereby keeping M_2-M_3 and M_3-M_4 simultaneously on. The circuit may therefore oscillate at $f_1/2$ (or injection-pulled by the clock).



Example 10.20 *(Continued)*

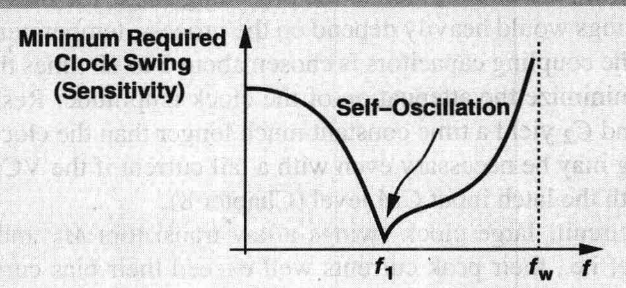


Figure 10.48 *Divider sensitivity plot.*

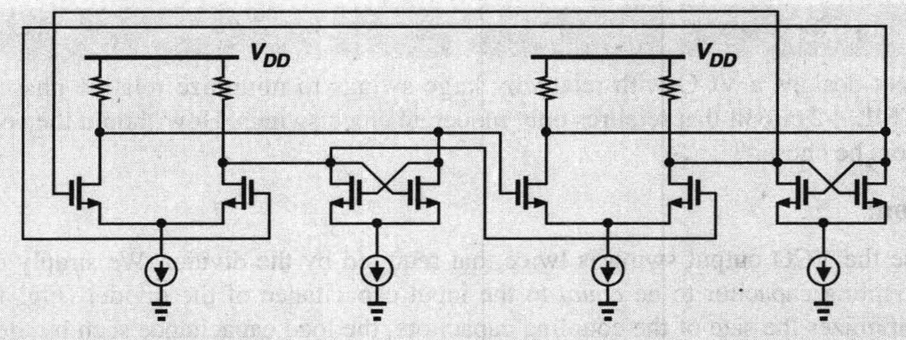


Figure 10.49 *Divide-by-2 circuit viewed as ring oscillator.*

The "self-oscillation" of the divider also proves helpful in the design process: if the choice of device dimensions does not allow self-oscillation, then the divider fails to operate properly. We thus first test the circuit with a zero clock swing to ensure that it oscillates.

The stacking of the differential and regenerative pairs atop the clocked pair in Figs. 10.43 and 10.47 does not lend itself to low supply voltages. This issue is alleviated by omitting the tail current source, but the bias currents of the circuit must still be defined accurately. Figure 10.50 shows an example [9], where the bias of the clocked pair is defined

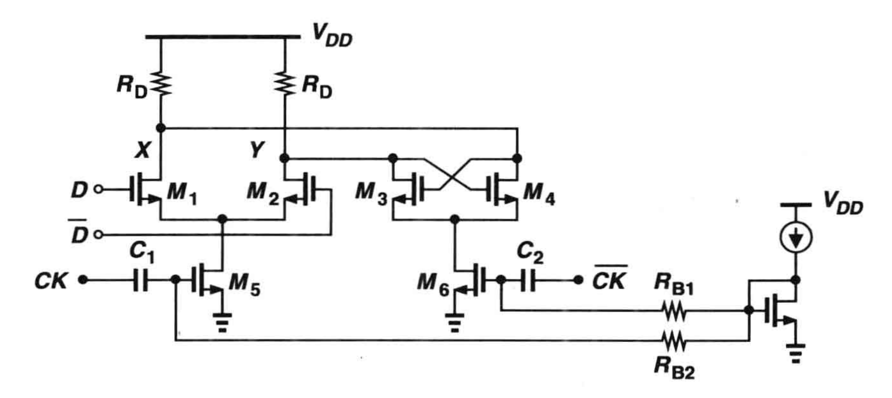


Figure 10.50 *Class-AB latch.*

by a current mirror and the clock is coupled capacitively. Without a current mirror, i.e., if the gates of M_5 and M_6 are directly tied to the preceding stage, the bias currents and hence the latch output swings would heavily depend on the process, temperature, and supply voltage. The value of the coupling capacitors is chosen about 5 to 10 times the gate capacitance of M_5 and M_6 to minimize the attenuation of the clock amplitude. Resistors R_{B1} and R_{B2} together with C_1 and C_2 yield a time constant much longer than the clock period. Note that capacitive coupling may be necessary even with a tail current if the VCO output CM level is incompatible with the latch input CM level (Chapter 8).

In the above circuit, large clock swings allow transistors M_5 and M_6 to operate in the class AB mode, i.e., their peak currents well exceed their bias current. This attribute improves the speed of the divider [9].

Example 10.21

A student designs a VCO with relatively large swings to minimize relative phase noise and a CML ÷2 circuit that requires only moderate clock swings. How should the coupling capacitors be chosen?

Solution:

Suppose the VCO output swing is twice that required by the divider. We simply choose each coupling capacitor to be *equal* to the input capacitance of the divider (Fig. 10.51). This minimizes the size of the coupling capacitors, the load capacitance seen by the VCO (*half* of the divider input capacitance), and the effect of divider input capacitance variation on the VCO.

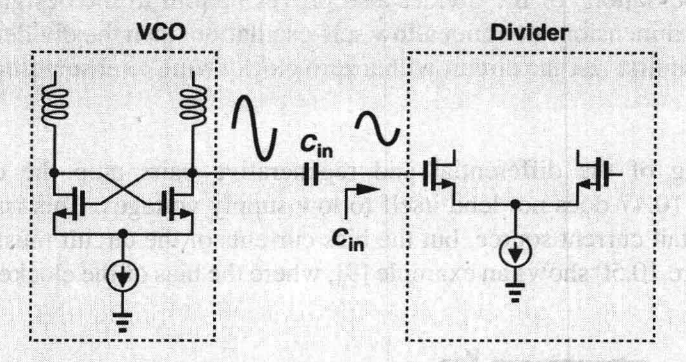


Figure 10.51 *Use of coupling capacitor equal to input capacitance of next stage.*

Recall from Chapter 4 that a VCO/÷2 circuit cascade proves useful in both generating the I and Q phases of the LO and avoiding injection pulling by the PA. This topology, however, dictates operation at twice the carrier frequency of interest. For the ÷2 stage to run at high frequencies, the speed of the D latch of Fig. 10.43 or 10.50 must be maximized. For example, inductive peaking raises the bandwidth at the output nodes [Fig. 10.52(a)]. From a small-signal perspective, we observe that the inductors rise in impedance at higher frequencies, allowing more of the currents produced by the transistors to flow through the capacitors and hence generate a larger output voltage. This behavior can be formulated with

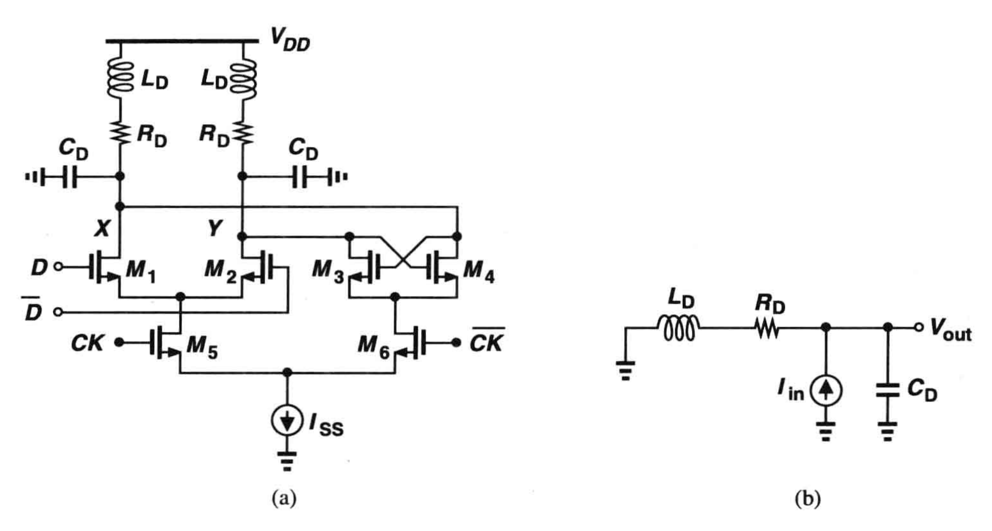


Figure 10.52 *(a) CML latch using inductive peaking, (b) equivalent circuit.*

the aid of the equivalent circuit shown in Fig. 10.52(b). We have

$$\frac{V_{out}}{I_{in}} = \frac{L_D s + R_D}{L_D C_D s^2 + R_D C_D s + 1}. \tag{10.50}$$

It is common to rewrite this transfer function as

$$\frac{V_{out}}{I_{in}} = \frac{s + 2\zeta\omega_n}{s^2 + 2\zeta\omega_n s + \omega_n^2} \cdot \frac{1}{C_D}, \tag{10.51}$$

where

$$\zeta = \frac{R_D}{2}\sqrt{\frac{C_D}{L_D}} \tag{10.52}$$

is the "damping factor" and

$$\omega_n = \frac{1}{\sqrt{L_D C_D}} \tag{10.53}$$

is the "natural frequency." To determine the -3-dB bandwidth, we equate the squared magnitude of Eq. (10.51) to $(1/2)(2\zeta/\omega_n)^2(1/C_D)^2$:

$$\frac{\omega_{-3dB}^2 + 4\zeta^2\omega_n^2}{(\omega_{-3dB}^2 - \omega_n^2)^2 + 4\zeta^2\omega_n^2\omega_{-3dB}^2} = \frac{2\zeta^2}{\omega_n^2}. \tag{10.54}$$

It follows that

$$\omega_{-3dB}^2 = \left[-2\zeta^2 + 1 + \frac{1}{4\zeta^2} + \sqrt{\left(-2\zeta^2 + 1 + \frac{1}{4\zeta^2}\right)^2 + 1}\right]\omega_n^2. \tag{10.55}$$

For example, noting that $\omega_n = 2\zeta/(R_D C_D)$, we obtain $\omega_{-3dB} \approx 1.8/(R_D C_D)$ if $\zeta = 1/\sqrt{2}$, i.e., the bandwidth increases by 80%. If L_D is increased further so that $\zeta = 1/\sqrt{3}$, then $\omega_{-3dB} = 1.85/(R_D C_D)$. On the other hand, a conservative value of $\zeta = 1$ yields $\omega_{-3dB} = 1.41/(R_D C_D)$.

Example 10.22

What is the minimum tolerable value of ζ if the frequency response must exhibit no peaking?

Solution:

Peaking occurs if the magnitude of the transfer function reaches a local maximum at some frequency. Taking the derivative of the magnitude squared of Eq. (10.51) with respect to ω and setting the result to zero, we have

$$\omega^4 + 8\zeta^2\omega_n^2\omega^2 + [4\zeta^2(4\zeta^2 - 2) - 1]\omega_n^4 = 0. \tag{10.56}$$

A solution exists if

$$-4\zeta^2 + \sqrt{8\zeta^2 + 1} \geq 0 \tag{10.57}$$

and hence if

$$\zeta \geq \sqrt{\frac{1+\sqrt{2}}{4}} \approx 0.78. \tag{10.58}$$

This bound on ζ translates to

$$\omega_{-3dB} \leq \frac{1.73}{R_D C_D}. \tag{10.59}$$

In practice, parasitics of on-chip inductors yield a bandwidth improvement less than the values predicted above. One may consider the Q of the inductor unimportant as R_D appears in series with L_D in Fig. 10.52. However, Q does play a role because of the finite parasitic *capacitance* of the inductor. For this reason, it is preferable to tie L_D to V_{DD} and R_D to the output node than L_D to the output node and R_D to V_{DD}; in the circuit of Fig. 10.52(a), about half of the distributed capacitance of L_D is absorbed by V_{DD}.[7] This also allows a symmetric inductor and hence a higher Q.

The topology depicted in Fig. 10.52(a) is called "shunt peaking" because the resistor-inductor branch appears in parallel with the output port. It is also possible to incorporate "series peaking," whereby the inductor is placed in series with the unwanted capacitance. Illustrated in Fig. 10.53, the circuit provides the following transfer function:

$$\frac{V_{out}}{I_{in}}(s) = \frac{R_D}{L_D C_D s^2 + R_D C_D s + 1}, \tag{10.60}$$

7. It can be proved that, in fact, 2/3 of the distributed capacitance is absorbed by the ac ground (Chapter 7).

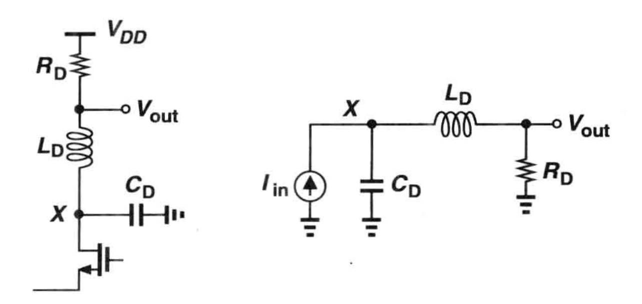


Figure 10.53 *Series peaking.*

which is similar to Eq. (10.50) but for a zero. The -3-dB bandwidth is computed as

$$\omega_{-3dB}^2 = [-(2\zeta^2 - 1) + \sqrt{(2\zeta^2 - 1)^2 + 1}]\omega_n^2, \tag{10.61}$$

where ζ and ω_n are given by (10.52) and (10.53), respectively. For example, if $\zeta = 1/\sqrt{2}$, then $\omega_{-3dB} = \omega_n = \sqrt{2}/(R_D C_D)$, i.e., series peaking increases the bandwidth by about 40%. The reader can prove that the frequency response exhibits peaking if $\zeta < 1/\sqrt{2}$. Note that V_X/I_{in} satisfies the shunt peaking transfer function of (10.50).

Example 10.23

Having understood shunt peaking intuitively, a student reasons that series peaking *degrades* the bandwidth because, at high frequencies, inductor L_D in Fig. 10.53 impedes the flow of current, forcing a larger fraction of I_{in} to flow through C_D. Since a smaller current flows though L_D and R_D, V_{out} falls at higher frequencies. Explain the flaw in this argument.

Solution:

Let us study the behavior of the circuit at $\omega_n = 1/\sqrt{L_D C_D}$. As shown in Fig. 10.54(a), the Thevenin equivalent of I_{in}, C_D, and L_D is constructed by noting that (a) the open-circuit output voltage is equal to $I_{in}/(C_D s)$, and (b) the output impedance (with I_{in} set to zero) is *zero* because C_D and L_D resonate at ω_n. It follows that $V_{out} = I_{in}/(C_D s)$ at $\omega = \omega_n$, i.e., as if the circuit consisted of only I_{in} and C_D [Fig. 10.54(b)]. Since I_{in} appears to flow entirely through C_D, it yields a larger magnitude for V_{out} than if it must split between C_D and R_D.

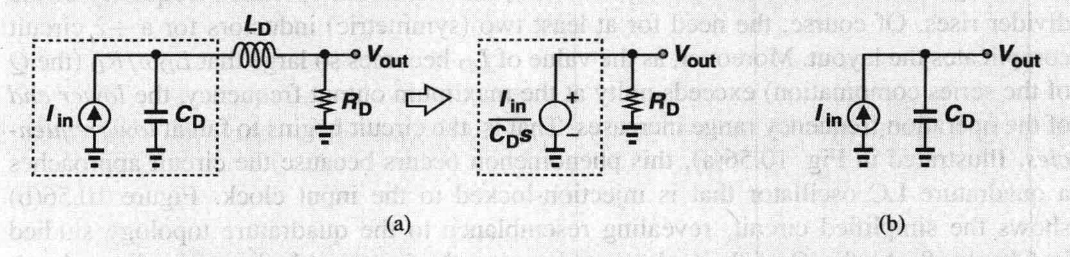


Figure 10.54 *(a) Use of Thevenin equivalent at resonance frequency, (b) simplified view.*

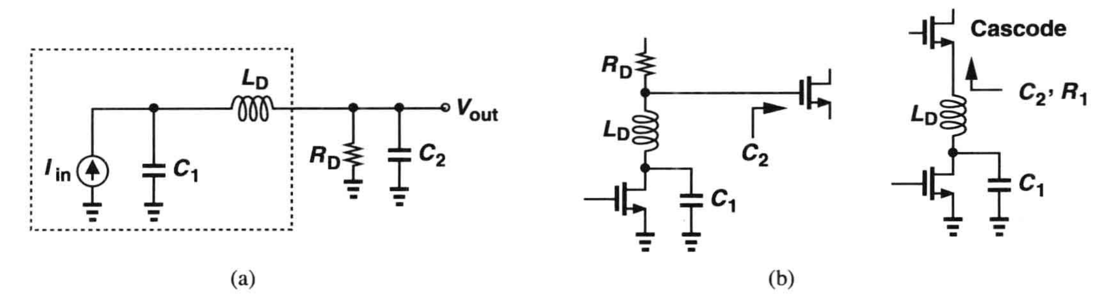


Figure 10.55 *(a) Series peaking circuit driving load capacitance C_2, (b) representative cases.*

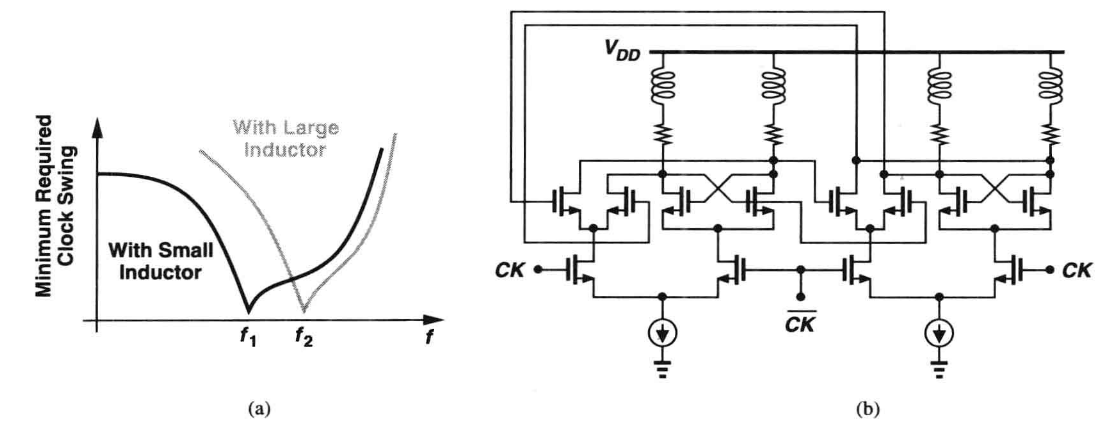


Figure 10.56 *(a) Sensitivity plots with small and large load inductors, (b) inductively-peaked latch viewed as quadrature oscillator.*

Circuits employing series peaking are generally more complex than the situation portrayed in Fig. 10.53. Specifically, the transistor generating I_{in} suffers from an output capacitance, which can be represented by C_D, but the next stage also exhibits an input capacitance, which is part of C_D in Fig. 10.52(a) but not included in Fig. 10.53. A more complete model is shown in Fig. 10.55(a) and two circuit examples employing series peaking are depicted in Fig. 10.55(b). The transfer function of the topology in Fig. 10.55(a) is of third order, making it difficult to compute the bandwidth, but simulations can be used to quantify the performance. Compared to shunt peaking, series peaking typically requires a smaller inductor value.

As L_D increases from zero in Fig. 10.52(a), the maximum operation frequency of the divider rises. Of course, the need for at least two (symmetric) inductors for a $\div 2$ circuit complicates the layout. Moreover, as the value of L_D becomes so large that $L_D\omega/R_D$ (the Q of the series combination) exceeds unity at the maximum output frequency, the *lower end* of the operation frequency range increases. That is, the circuit begins to fail at *low frequencies*. Illustrated in Fig. 10.56(a), this phenomenon occurs because the circuit approaches a quadrature LC oscillator that is injection-locked to the input clock. Figure 10.56(b) shows the simplified circuit, revealing resemblance to the quadrature topology studied in Chapter 8. As the Q of the tank exceeds unity, the injection lock range of the circuit becomes narrower.

CML dividers have reached very high speeds in deep-submicron CMOS technologies. For example, the use of inductive peaking and class-AB operation has afforded a maximum clock frequency of 96 GHz for a ÷2 circuit [10].

True Single-Phase Clocking Another logic style often employed in divider design is "true single-phase clocking" (TSPC) [11]. Figure 10.57(a) shows a TSPC flipflop. Incorporating dynamic logic, the circuit operates as follows. When CK is high, the first stage operates as an inverter, impressing $\overline{D}$ at A and E. When CK goes low, the first stage is disabled and the second stage becomes transparent, "writing" $\overline{A}$ at B and C and hence making Q equal to A. The logical high at E and the logical low at B are degraded, but the levels at A and C ensure proper operation of the circuit.

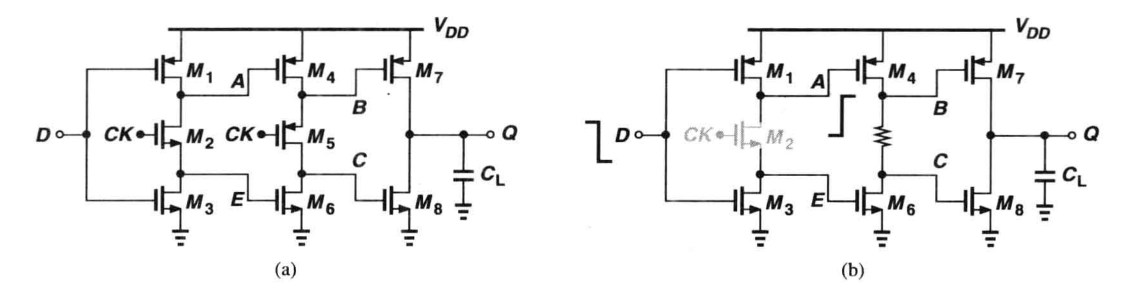


Figure 10.57 *(a) TSPC flipflop, (b) response to a rising input transition when CK is low.*

Does the clock completely disable each of the first two stages? Suppose CK is low. If D goes from low to high, A remains constant. On the other hand, if D falls [Fig. 10.57(b)], A rises, but the state at B does not change because M_4 turns off and M_6 remains off. (For a state to change, a transistor must turn *on*.) Now suppose CK is high, keeping M_2 on and M_5 off. Does a change in D propagate to Q? For example, if Q is high, and D rises, then A falls and B rises, turning M_7 *off*. Thus, Q remains unchanged.

Since the flipflop of Fig. 10.57 contains one inversion, it can serve as a ÷2 circuit if Q is tied to D. An alternative ÷2 TSPC circuit is shown in Fig. 10.58 [11]. This topology achieves relatively high speeds with low power dissipation, but, unlike CML dividers, it requires rail-to-rail clock swings for proper operation. Moreover, it does not provide quadrature outputs. Note that (a) the circuit consumes no static power,[8] and (b) as a dynamic

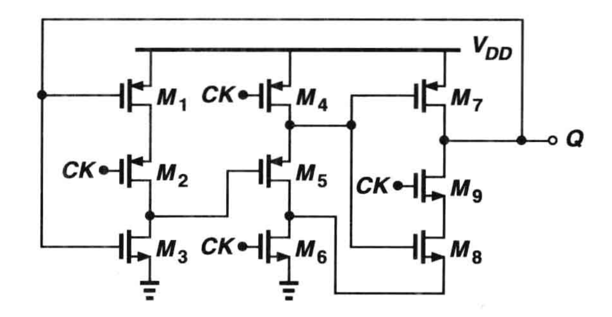


Figure 10.58 *TSPC divide-by-2 circuit.*

8. Except for the subthreshold leakage of the transistors.

logic topology, the divider fails at very *low* clock frequencies due to the leakage of the transistors. For example, if a synthesizer is designed with a reference frequency of 1 MHz, then the last few stages of the program counter in Fig. 10.24 must operate at a few megahertz, possibly failing to retain states stored on transistor capacitances.

The TSPC FF of Fig. 10.57 can readily incorporate logic at its input. For example, a NAND gate can be merged with the master latch as shown in Fig. 10.59. Thus circuits such as the ÷3 stage of Fig. 10.31 can be realized by TSPC logic as well. In the design of TSPC circuits, one observes that wider clocked devices [e.g., M_2 and M_5 in Fig. 10.57(a)] raises the maximum speed, but at the cost of loading the preceding stage, e.g., the VCO.

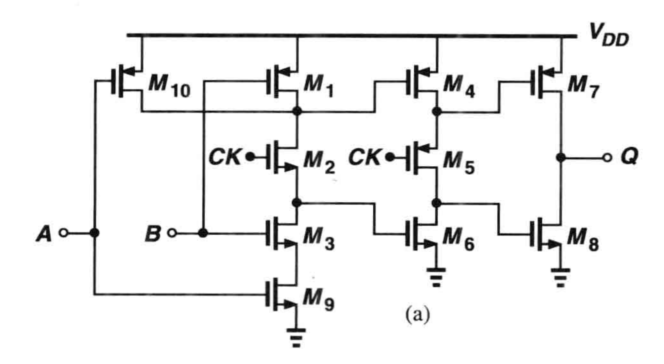


Figure 10.59 *TSPC FF incorporating a NAND gate.*

A variant of TSPC logic that achieves higher speeds is depicted in Fig. 10.60 [12]. Here, the first stage operates as the master D latch and the last two as the slave D latch. The slave latch is designed as "ratioed" logic, i.e., both NMOS devices are strong enough to pull down *B* and *Q* even if M_4 or M_6 is on. When *CK* is high, the first stage reduces to an inverter, the second stage forces a ZERO at *B*, and the third stage is in the store mode. When *CK* goes down, *B* remains low if *A* is high, or it rises if *A* is low, with *Q* tracking *B* because M_7 and M_6 act as a ratioed inverter.

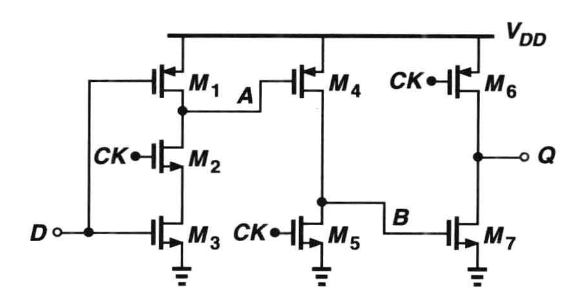


Figure 10.60 *TSPC circuit using ratioed logic.*

Example 10.24

The first stage in Fig. 10.60 is not completely disabled when *CK* is low. Explain what happens if *D* changes in this mode.

Example 10.24 *(Continued)*

Solution:

If D goes from low to high, A does not change. If D falls, A rises, but since M_4 turns *off*, it cannot change the state at B. Thus, D does not alter the state stored by the slave latch.

The second and third stages in the circuit of Fig. 10.60 consume static power when their clocked transistor fights the input device. At high speeds, however, the dynamic power dominates, making this drawback less objectionable. In a typical design (with PMOS mobility about half of NMOS mobility), all transistors in the circuit can have equal dimensions, except for W_5, which must be two to three times the other transistor widths to maximize the speed. The input can also incorporate logic in a manner similar to that in Fig. 10.59. This technique allows $\div 2$ speeds around 10 GHz and $\div 3$ speeds around 6 GHz in 65-nm CMOS technology. We incorporate this logic style in the design of a $\div 3/4$ circuit in Chapter 13.

The TSPC circuit and its variants operate with rail-to-rail swings but do not provide differential or quadrature outputs. A complementary logic style resolving this issue is described in Chapter 13.

10.6.5 Miller Divider

As explained in Section 10.6.4, CML dividers achieve a high speed by virtue of current steering and moderate voltage swings. If the required speed exceeds that provided by CML circuits, one can consider the "Miller divider" [13], also known as the "dynamic divider."[9] Depicted in Fig. 10.61(a) and providing a divide ratio of 2, the Miller topology consists of a mixer and a low-pass filter, with the LPF output fed back to the mixer. If the circuit operates properly, $f_{out} = f_{in}/2$, yielding two components, $3f_{in}/2$ and $f_{in}/2$, at node X. The former is attenuated by the LPF, and the latter circulates around the loop. In other words, correct operation requires that the loop gain for the former component be sufficiently small and that for the latter exceed unity.

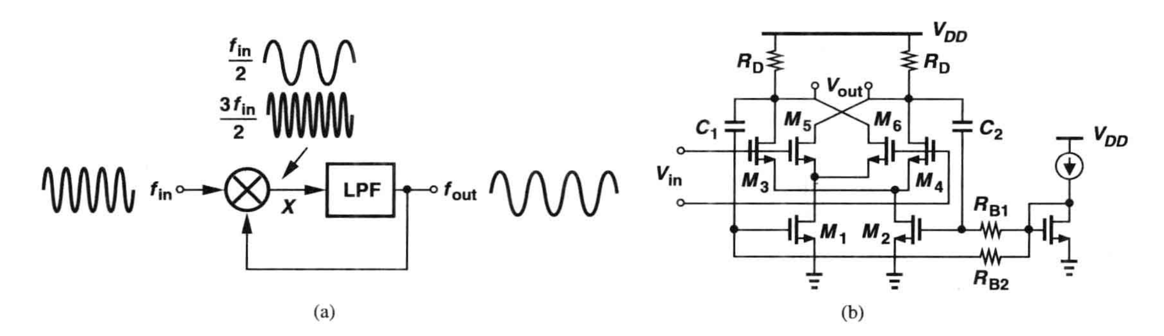


Figure 10.61 *(a) Miller divider, (b) realization.*

9. But this terminology must not be confused with dynamic logic.

The Miller divider can achieve high speeds for two reasons: (1) the low-pass behavior can simply be due to the intrinsic time constant at the output node of the mixer, and (2) the circuit does not rely on latching and hence fails more gradually than flipflops as the input frequency increases. Note, however, that the divider loop requires some cycles to reach steady state, i.e., it does not divide correctly instantaneously.

Figure 10.61(b) shows an example of the Miller divider realization. A double-balanced mixer senses the input at its LO port, with C_1 and C_2 returning the output to the RF port.[10] The loop gain at mid-band frequencies is equal to $(2/\pi)g_{m1,2}R_D$ (Chapter 6) and must remain above unity. The maximum speed of the circuit is roughly given by the frequency at which the loop gain falls to unity. Note that R_D and the total capacitance at the output nodes define the corner of the LPF. Of course, at high frequencies, the roll-off due to the pole at the drains of M_1 and M_2 may also limit the speed.

Example 10.25

Is it possible to construct a Miller divider by returning the output to the LO port of the mixer?

Solution:

Shown in Fig. 10.62, such a topology senses the input at the RF port of the mixer. (Strangely enough, M_3 and M_4 now appear as diode-connected devices.) We will see below that this circuit fails to divide.

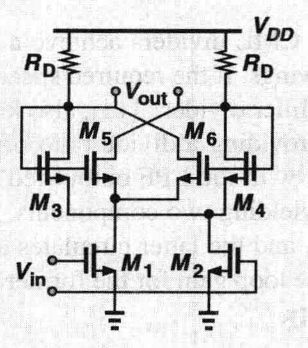


Figure 10.62 *Miller divider with feedback to switching quad.*

The reader may wonder why we said above that the component at $3f_{in}/2$ must be sufficiently small. After all, this component is merely the third harmonic of the desired output and would seem to only *sharpen* the output edges. However, this harmonic in fact creates a finite *lower* bound on the divider operation frequency. As f_{in} decreases and the component at $3f_{in}/2$ falls *below* the corner frequency of the LPF, the circuit fails to divide.

10. Capacitive coupling ensures that M_1 and M_2 can operate in saturation.

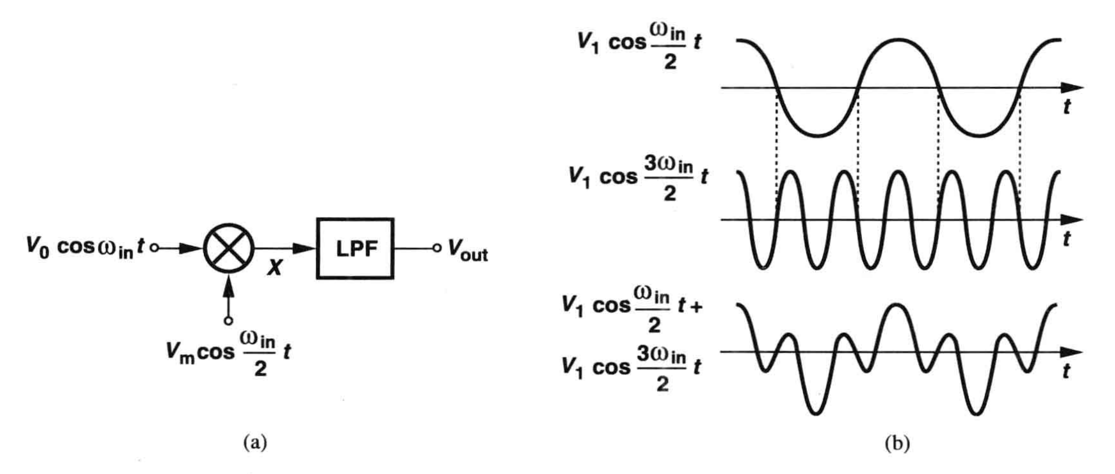


Figure 10.63 *(a) Open-loop equivalent circuit of Miller divider, (b) circuit's waveforms.*

To understand this point, let us open the loop as shown in Fig. 10.63(a) and write the output of the mixer as

$$V_X(t) = \alpha(V_0 \cos \omega_{in} t)\left(V_m \cos \frac{\omega_{in} t}{2}\right) \tag{10.62}$$

$$= \frac{\alpha V_0 V_m}{2}\left(\cos \frac{\omega_{in} t}{2} + \cos \frac{3\omega_{in} t}{2}\right), \tag{10.63}$$

where α is related to the mixer conversion gain. Illustrated in Fig. 10.63(b), this sum exhibits additional zero crossings, prohibiting frequency division if traveling through the LPF unchanged [14]. Thus, the third harmonic must be attenuated—at least by a factor of three [14]—to avoid the additional zero crossings.

Example 10.26

Does the arrangement shown in Fig. 10.64 operate as a divider?

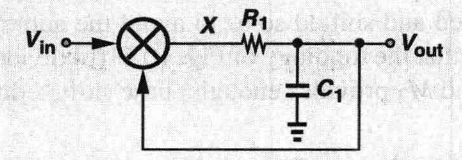


Figure 10.64 *Miller divider using first-order low-pass filter.*

(Continues)

Example 10.26 *(Continued)*

Solution:

Since the voltage drop across R_1 is equal to $R_1C_1dV_{out}/dt$, we have $V_X = R_1C_1dV_{out}/dt + V_{out}$. Also, $V_X = \alpha V_{in}V_{out}$. If $V_{in} = V_0 \cos \omega_{in}t$, then

$$R_1C_1\frac{dV_{out}}{dt} + V_{out} = \alpha(V_0 \cos \omega_{in}t)V_{out}. \tag{10.64}$$

It follows that

$$R_1C_1\frac{dV_{out}}{V_{out}} = (\alpha V_0 \cos \omega_{in}t - 1)dt. \tag{10.65}$$

We integrate the left-hand side from V_{out0} (initial condition at the output) to V_{out} and the right-hand side from 0 to t:

$$R_1C_1 \ln\frac{V_{out}}{V_{out0}} = \frac{1}{\omega_{in}}\alpha V_0 \sin \omega_{in}t - t. \tag{10.66}$$

Thus,

$$V_{out}(t) = V_{out0} \exp\left(\frac{-t}{R_1C_1} + \frac{\alpha V_0}{R_1C_1\omega_{in}} \sin \omega_{in}t\right). \tag{10.67}$$

Interestingly, the exponential term drives the output to zero regardless of the values of α or ω_{in} [14]. The circuit fails because a one-pole filter does not sufficiently attenuate the third harmonic with respect to the first harmonic. An important corollary of this analysis is that the topology of Fig. 10.62 cannot divide: the single-pole loop follows Eq. (10.67) and does not adequately suppress the third harmonic at the output.

To avoid the additional zero crossings shown in Fig. 10.63(b), it is also possible to introduce *phase shift* in $\cos(\omega_{in}t/2)$ and/or $\cos(3\omega_{in}t/2)$. For example, if $\cos(3\omega_{in}t/2)$ is attenuated by a factor of 2 but shifted by 45°, then the two components add up to the waveform shown in Fig. 10.65. In other words, the Miller divider operates properly if the third harmonic is attenuated and shifted so as to avoid the additional zero crossings [14]. This observation suggests that the topology of Fig. 10.61(b) divides successfully only if the pole at the drains of M_1 and M_2 provides enough phase shift, a difficult condition.

Miller Divider with Inductive Load The topology of Fig. 10.61(b) suffers from the same gain-headroom trade-offs as those described for active mixers in Chapter 6. The limited voltage drop across the load resistors makes it difficult to achieve a high conversion gain. Wider input and LO transistors alleviate this issue but at the cost of speed. If the load resistors are replaced with inductors, the gain-headroom and gain-speed trade-offs are greatly relaxed, but the lower end of the frequency range rises. Also, the inductor complicates the

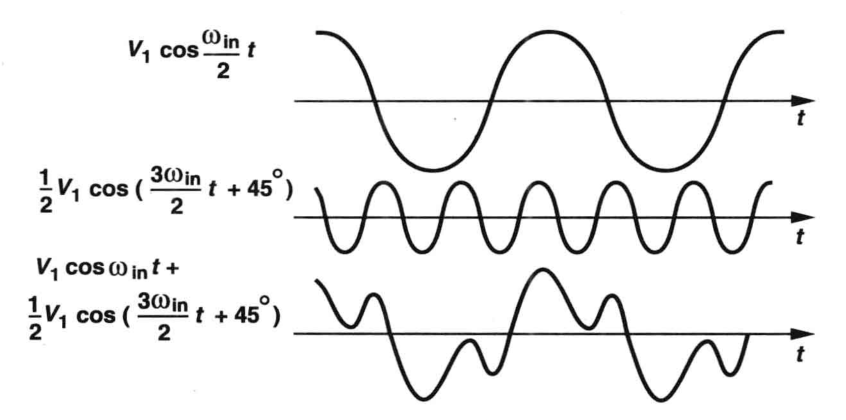


Figure 10.65 *Proper Miller divider operation if third harmonic is shifted by 45° and halved in amplitude.*

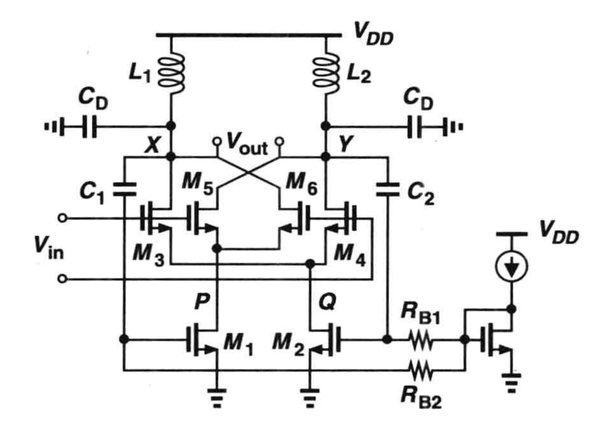


Figure 10.66 *Miller divider using inductive loads.*

layout. Figure 10.66 shows a Miller divider using inductive loads. Since the tanks significantly suppress the third harmonic of the desired output, this circuit proves more robust than the topology of Fig. 10.61(b) [14].

The design of this circuit proceeds as follows. We assume a certain tolerable capacitance at the LO port and a certain load capacitance at the output node. The width of M_3-M_6 is chosen according to the former, and the LO common-mode level is preferably chosen equal to V_{DD}, leaving maximum headroom for M_1 and M_2. Inductors L_1 and L_2 must resonate with the total capacitance at X and Y at about *half* of the input "mid-band" frequency. For example, if we wish to accommodate an input range of 40 GHz to 50 GHz, we may choose a resonance frequency around 22.5 GHz. In this step, the capacitance of M_1 and M_2 is unknown, requiring a reasonable guess and possible iterations. With the value of L_1 and L_2 known, their equivalent parallel resistance, R_p, must provide enough gain along with M_1 and M_2. [Recall that the mixer has a conversion gain of $(2/\pi)g_{m1,2}R_p$.] The width and bias current of M_1 and M_2 are therefore chosen so as to maximize the gain. Of course, an excessively high bias current results in a low value for V_P and V_Q, driving M_1 and M_2 into the triode region. Some optimization is thus necessary. It can be shown that the input

frequency range across which the circuit operates properly is given by

$$\Delta\omega = \frac{2\omega_0}{Q}\left(\frac{2}{\pi}g_{m1,2}R_p\right)^2, \tag{10.68}$$

where ω_0 and Q denote the tank resonance frequency and quality factor, respectively [14].

Example 10.27

Does the circuit of Fig. 10.62 operate as a divider if the load resistors are replaced with inductors?

Solution:

Depicted in Fig. 10.67(a), such an arrangement in fact resembles an *oscillator*. Redrawing the circuit as shown in Fig. 10.67(b), we note M_5 and M_6 act as a cross-coupled pair and M_3 and M_4 as diode-connected devices. In other words, the oscillator consisting of M_5-M_6 and L_1-L_2 is heavily loaded by M_3-M_4, failing to oscillate (unless the Q of the tank is infinite or M_3 and M_4 are weaker than M_5 and M_6). This configuration does operate as a divider but across a narrower frequency range than does the topology of Fig. 10.66.

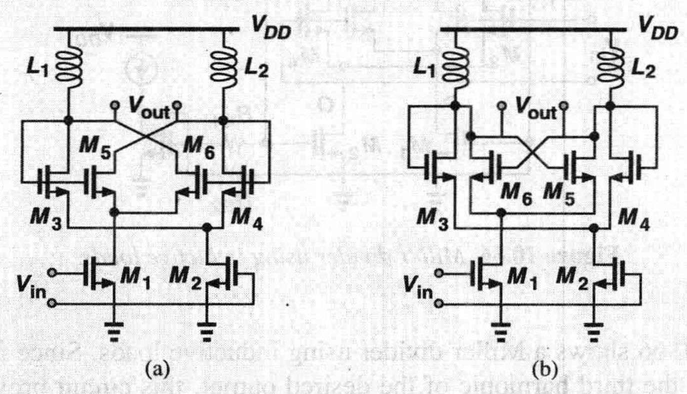


Figure 10.67 *(a) Inductively-loaded Miller divider with feedback to the switching quad, (b) alternative drawing of the circuit.*

It is possible to construct a Miller divider using *passive* mixers. Figure 10.68 depicts an example, where M_1-M_4 constitute a passive mixer and M_5-M_6 an amplifier [16]. Since the output CM level is near V_{DD}, the feedback path incorporates capacitive coupling, allowing the sources and drains of M_1-M_4 to remain about 0.4 V above ground. (As explained in Chapter 6, the LO CM level must still be near V_{DD} to provide sufficient overdrive for M_1-M_4.) The cross-coupled pair M_7-M_8 can be added to increase the gain by virtue of its negative resistance. If excessively strong, however, this pair oscillates with the tanks, leading to a narrow frequency range (Section 10.6.6).

While achieving speeds exceeding 100 GHz in CMOS technology [16], the Miller divider does not provide quadrature outputs, a drawback with respect to RF transceivers that operate the oscillator at twice the carrier frequency and employ a divider to generate the I and Q phases.

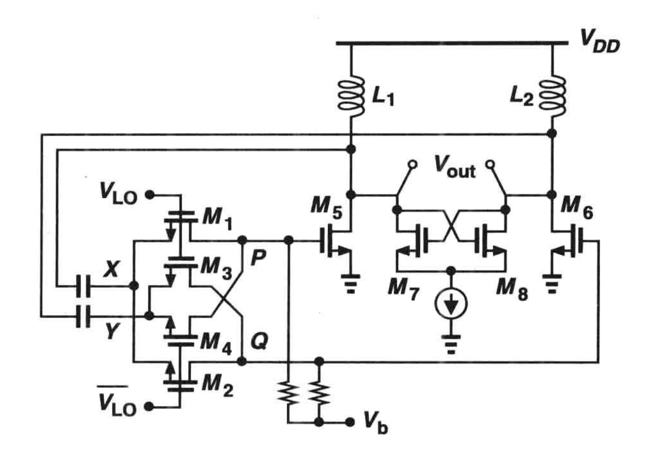


Figure 10.68 *Miller divider using passive mixer.*

Miller Divider with Other Moduli In his original paper, Miller also contemplates the use of dividers *within* the feedback loop of his topology so as to produce moduli other than 2. Shown in Fig. 10.69 is an example, where a $\div N$ circuit in the feedback path creates $f_b = f_{out}/N$, yielding $f_{in} \pm f_{out}/N$ at X. If the sum is suppressed by the LPF, then $f_{out} = f_{in} - f_{out}/N$ and hence

$$f_{out} = \frac{N}{N+1} f_{in}. \tag{10.69}$$

Also,

$$f_b = \frac{1}{N+1} f_{in}. \tag{10.70}$$

For example, if $N = 2$, the input frequency is divided by 1.5 and 3, two moduli that are difficult to obtain at high speeds by means of flipflop-based dividers.

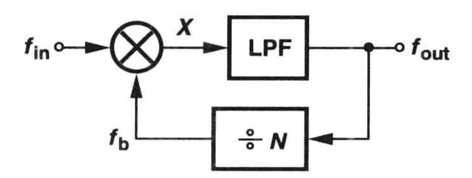


Figure 10.69 *Miller divider having another divider stage in feedback.*

An important issue in the topology of Fig. 10.69 is that the sum component at X comes closer to the difference component as N increases, dictating a sharper LPF roll-off. In the above example, these two components lie at $4f_{in}/3$ and $2f_{in}/3$, respectively, i.e., only one octave apart. Consequently, the circuit suffers from a more limited frequency range.

Another critical issue in the Miller divider of Fig. 10.69 relates to the port-to-port feedthroughs of the mixer. The following example illustrates this point.

Example 10.28

Assume $N = 2$ in Fig. 10.69 and study the effect of feedthrough from each input port of the mixer to its output.

(Continues)

Example 10.28 *(Continued)*

Solution:

Figure 10.70(a) shows the circuit. The feedthrough from the main input to node X produces a spur at f_{in}. Similarly, the feedthrough from Y to X creates a component at $f_{in}/3$. The output therefore contains two spurs around the desired frequency [Fig. 10.70(b)]. Interestingly, the signal at Y exhibits no spurs: as the spectrum of Fig. 10.70(b) travels through the divider, the main frequency component is divided while the spurs maintain their spacing with respect to the carrier (Chapter 9). Shown in Fig. 10.70(c), the spectrum at Y contains only harmonics and a dc offset. The reader can prove that these results are valid for any value of N.

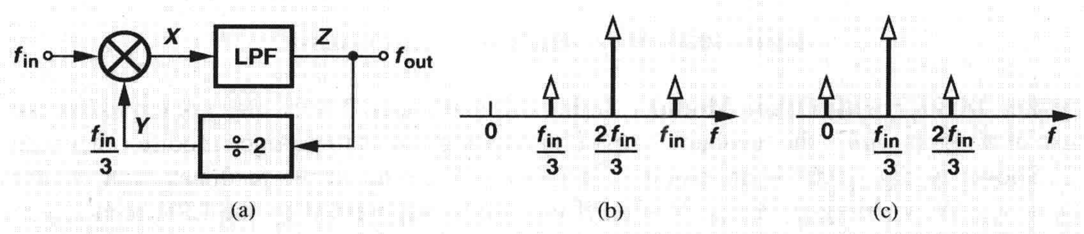


Figure 10.70 *(a) Miller divider with a ÷2 stage in feedback, (b) output spectrum, (c) spectrum at Y.*

Does the original topology of Fig. 10.61(a) also suffer from spurs? In Problem 10.16, we prove that it does not.

The Miller divider frequency range can be extended through the use of a single-sideband mixer. Illustrated in Fig. 10.71(a), the idea is to suppress the sum component by SSB mixing rather than filtering, thereby avoiding the problem of additional zero crossings depicted in Fig. 10.63. In the absence of the sum component, the circuit divides properly at arbitrarily low frequencies. Unfortunately, however, this approach requires a broadband 90° phase shift, a very difficult design.

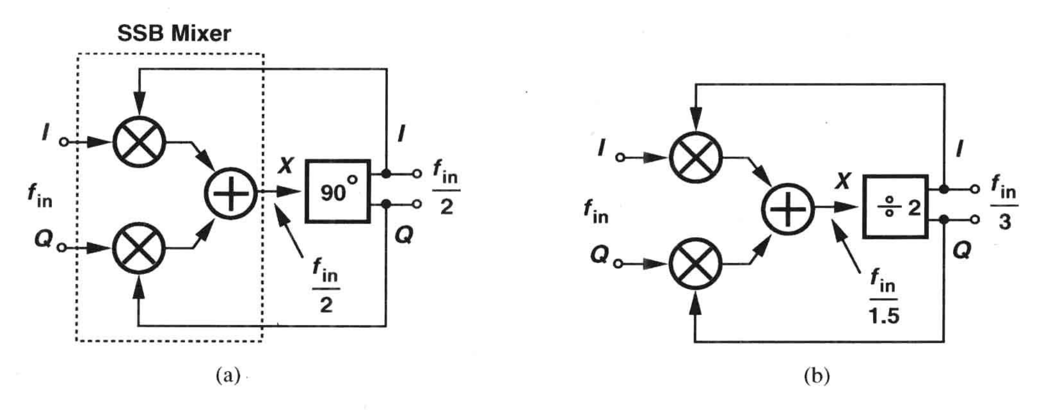


Figure 10.71 *(a) Miller divider using SSB mixer, (b) divide-by-3 realization.*

Nonetheless, the use of SSB mixing does prove useful if the loop contains a divider that generates quadrature outputs [17]. Shown in Fig. 10.71(b) is an example employing a ÷2 circuit and generating $f_{in}/3$ at the output [17]. This topology achieves a wide frequency range *and* generates quadrature outputs, a useful property for multiband applications. By contrast, the flipflop-based ÷3 circuit in Fig. 10.28 does not provide quadrature outputs. We note from Example 10.28 that the ÷1.5 output available at X exhibits spurs and may not be suited to stringent systems.

The principal drawback of the circuit of Fig. 10.71(b) and its variants is that it requires quadrature LO phases. As explained in Chapter 8, quadrature oscillators exhibit a higher phase noise and two possible modes.

10.6.6 Injection-Locked Dividers

Another class of dividers is based on oscillators that are injection-locked to a harmonic of their oscillation frequency [15]. To understand this principle, let us return to the Miller divider of Fig. 10.68 and assume the cross-coupled pair is strong enough to produce oscillation. Transistors M_5 and M_6 can now be viewed as devices that *couple* the mixer output to the oscillator.[11] The overall loop can thus be modeled as shown in Fig. 10.72, where the connection between X and the oscillator denotes *injection* rather than frequency control. If the loop operates properly, then $f_{out} = f_{in}/2$, yielding both $f_{in}/2$ and $3f_{in}/2$ at X. The former couples to and locks the oscillator while the latter is suppressed by the selectivity of the oscillator. If f_{in} varies across a certain "lock range," the oscillator remains injection-locked to the $f_{out} - f_{in}$ component at node X. On the other hand, if f_{in} falls outside the lock range, the oscillator is injection-pulled, thus producing a corrupted output.

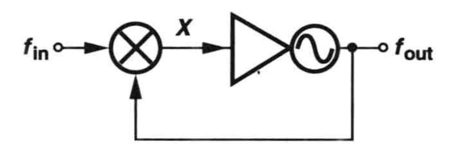


Figure 10.72 *Injection-locked divider.*

Example 10.29

Determine the divide ratio of the topology shown in Fig. 10.73 if the oscillator remains locked.

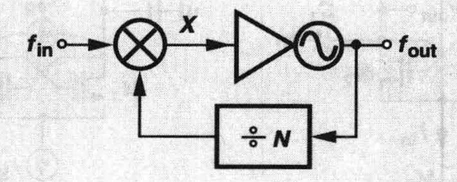


Figure 10.73 *Injection-locked divider having another divider in feedback.*

(Continues)

11. In a manner similar to the coupling mechanism in quadrature oscillators (Chapter 8).

Example 10.29 *(Continued)*

Solution:

The mixer yields two components at node X, namely, $f_{in} - f_{out}/N$ and $f_{in} + f_{out}/N$. If the oscillator locks to the former, then $f_{in} - f_{out}/N = f_{out}$ and hence

$$f_{out} = \frac{N}{N+1} f_{in}. \tag{10.71}$$

Similarly, if the oscillator locks to the latter, then

$$f_{out} = \frac{N}{N-1} f_{in}. \tag{10.72}$$

The oscillator lock range must therefore be narrow enough to lock to only one of the two components.

The reader may wonder if the injection-locked divider (ILD) of Fig. 10.72 is any different from the Miller loop of Fig. 10.61(a). The fundamental difference between the two is that the former oscillates even with the input amplitude set to zero, whereas the latter does not. For this reason, ILDs generally exhibit a narrower operation frequency range than Miller dividers.

Let us now implement an ILD. While the topology of Fig. 10.72 serves as a candidate, even a simpler arrangement is conceived if we recognize that the cross-coupled pair within an oscillator can also operate as a *mixer*. Indeed, we utilized this property to study the effect of the tail noise current on the phase noise in Chapter 8. The loop shown in Fig. 10.72 thus reduces to a single cross-coupled pair providing a negative resistance at its drains and a mixing input at its tail node. Figure 10.74(a) depicts the result: I_{in} $(= g_{m3}V_{in})$ is commutated by M_1 and M_2 and hence translated to $f_{out} \pm f_{in}$ as it emerges at the drains of these transistors. The circuit can be equivalently viewed as shown in Fig. 10.74(b), where

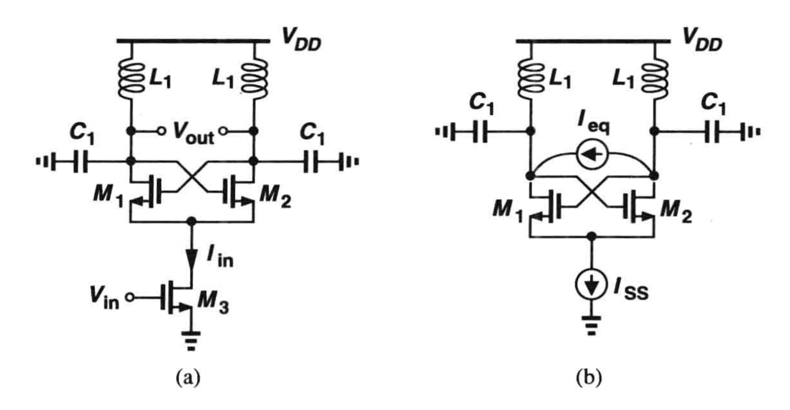


Figure 10.74 *(a) Example of injection-locked divider, (b) equivalent view.*

I_{eq} represents the current components at $f_{out} \pm f_{in}$, with the amplitude of each component given by $2/\pi$ times the amplitude of I_{in}.[12]

Since the sum component is greatly attenuated by the oscillator, we can consider only the difference component as the input to the oscillator. From Chapter 8, the *two-sided* injection lock range is given by $(\omega_0/Q)(I_{inj}/I_{osc})$, where $I_{inj} = (2/\pi)I_{in}$. Thus, the *output* frequency range across which the circuit remains locked is given by [18]

$$\Delta\omega_{out} = \frac{\omega_0}{Q}\left(\frac{2}{\pi}\frac{I_{in}}{I_{osc}}\right). \tag{10.73}$$

The *input* lock range is twice this value:

$$\Delta\omega_{in} = \frac{\omega_0}{Q}\left(\frac{4}{\pi}\frac{I_{in}}{I_{osc}}\right). \tag{10.74}$$

As explained in Chapter 8, the phase noise of an injection-locked oscillator approaches that of the unlocked circuit if the oscillator is locked near the edge of the lock range. Equation (10.74) therefore points to a trade-off between the lock range and the phase noise: as Q is lowered, the former widens but the latter degrades.

As mentioned in Section 10.6.4, flipflop-based dividers can also be considered injection-locked ring oscillators. Achieving a much wider lock range than their LC oscillator counterparts, these dividers exhibit a low phase noise by virtue of strong locking to the input. It is important to bear in mind that LC oscillators exhibit injection-locking dynamics that require a settling time commensurate with their Q. That is, such dividers do not begin to operate correctly instantaneously.

It is also critical to note that, in a PLL environment, the divider lock range must exceed the VCO tuning range. During the lock transient, the VCO frequency swings up or down and, if the divider fails at any VCO frequency, the PLL may simply not lock.

10.6.7 Divider Delay and Phase Noise

A divider incorporating asynchronous logic may experience a significant delay from the input to the output. For example, in the pulse swallow divider of Fig. 10.24, all three counters contribute delay: on one edge of the main input, the prescaler incurs some delay before it produces a transition at node A, which must then propagate through the program counter to reach the output.

What is the effect of divider delay on an integer-N synthesizer? The transfer function of a stage having a constant delay of ΔT is given by $\exp(-\Delta T \cdot s)$, yielding an overall open-loop transfer function of

$$H_{open}(s) = \frac{I_P}{2\pi}\left(R_P + \frac{1}{C_1 s}\right)\frac{K_{VCO}}{NS}e^{-\Delta T \cdot s}. \tag{10.75}$$

If the delay is small with respect to the time scales of interest, we can write $\exp(-\Delta T \cdot s) \approx 1 - \Delta T \cdot s$, recognizing that the delay results in a zero in the *right-half* plane. Such a zero

12. We assume M_1 and M_2 experience abrupt switching.

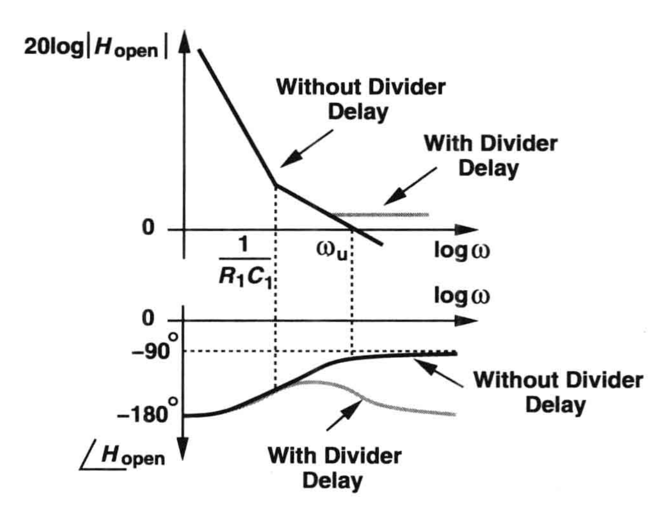


Figure 10.75 *Effect of divider delay on PLL phase margin.*

contributes *negative* phase shift, $-\tan^{-1}(\Delta T \cdot \omega)$, making the loop less stable. Plotted in Fig. 10.75 is the open-loop frequency response in this case, revealing that the zero has two undesirable effects: it flattens the gain, pushing the gain crossover frequency to *higher* values (in principle, infinity), and it bends the phase profile downward. Thus, this zero must remain well above the original unity-gain bandwidth of the loop, e.g.,

$$\frac{1}{\Delta T} \approx 5\omega_u \tag{10.76}$$

$$\approx 5(2\zeta^2 + \sqrt{4\zeta^4 + 1})\omega_n^2. \tag{10.77}$$

In most practical designs, the divider delay satisfies the above condition, thus negligibly affecting the loop dynamics. Otherwise, the divider must employ more synchronous logic to reduce the delay.

The divider phase noise may also prove troublesome. As shown in Fig. 10.76, the output phase noise of the divider, $\phi_{n,div}$, *directly* adds to the input phase noise, $\phi_{n,in}$, experiencing the same low-pass response as it propagates to ϕ_{out}. In other words, $\phi_{n,div}$ is also multiplied by a factor of N within the loop bandwidth. Thus, for the divider to contribute negligible phase noise, we must have

$$\phi_{n,div} \ll \phi_{n,in}. \tag{10.78}$$

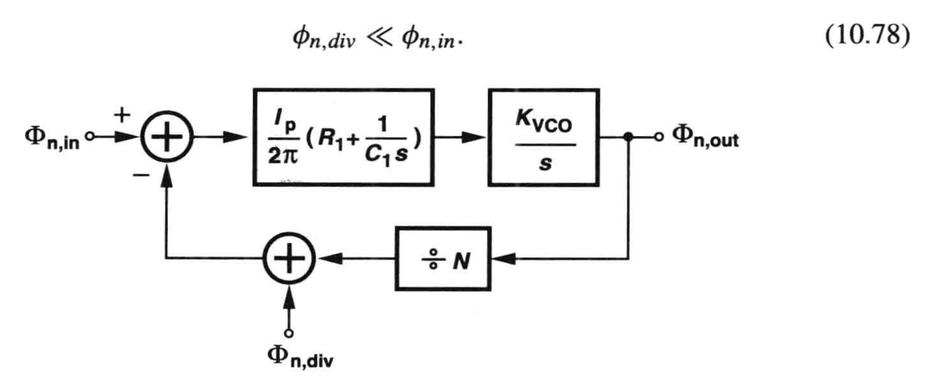
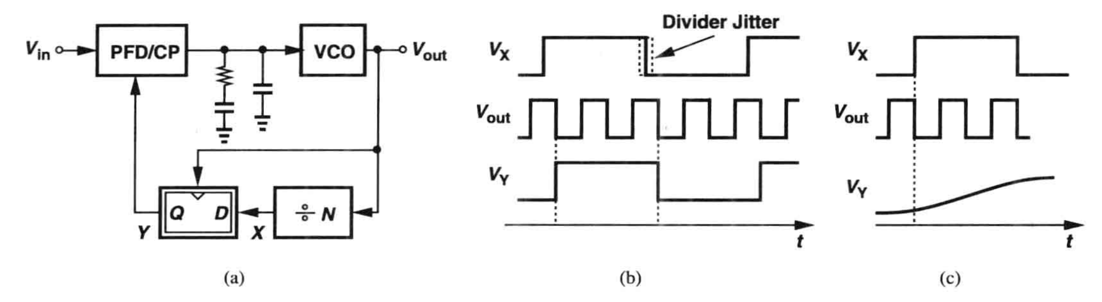


Figure 10.76 *Effect of divider phase noise on PLL.*



(a) (b) (c)

Figure 10.77 *(a) Use of retiming FF to remove divider phase noise, (b) waveforms showing retiming operation, (c) problem of metastability.*

In narrowband synthesizers, of course, the output phase noise is dominated by that of the VCO, making $\phi_{n,in}$ and $\phi_{n,div}$ less critical. Nonetheless, if the divider phase noise is significant, a retiming flipflop can be used to suppress its effect. Illustrated in Fig. 10.77(a), the idea is to sample the divider output by the VCO waveform, thus presenting the edges to the PFD only at the VCO transitions. As depicted in Fig. 10.77(b), V_Y changes only when the VCO output changes, avoiding the jitter (phase noise) in V_X (if the jitter is less than one cycle of V_{out}). In essence, the retiming operation bypasses the phase noise accumulated in the divider chain.

Example 10.30

Compare the output phase noise of the above circuit with that of a similar loop that employs noiseless dividers and no retiming flipflop. Consider only the input phase noise.

Solution:

The phase noise is similar. Invoking the time-domain view, we note that a (slow) displacement of the input edges by ΔT seconds still requires that the edges at Y be displaced by ΔT, which is possible only if the VCO edges are shifted by the same amount.

Example 10.31

Does the retiming operation in Fig. 10.77(a) remove the effect of the divider delay?

Solution:

No, it does not. An edge entering the divider still takes a certain amount of time before it appears at X and hence at Y. In fact, Fig. 10.77(b) indicates that V_Y is delayed with respect to V_X by at most one VCO cycle. That is, the overall feedback delay is slightly *longer* in this case.

A concern in the use of the retiming FF in Fig. 10.77(a) arises if the VCO output edge occurs close to the transition at node X. Under this condition, the FF becomes "metastable," i.e., it takes a long time to produce a well-defined logical level. Depicted in Fig. 10.77(c),

this effect results in a distorted transition at node Y, confusing the PFD. It is therefore essential to guarantee by design that the sampling edges of the VCO remain safely away from the transitions at node X. If the divider delay varies by as much as one VCO period with process and temperature, then it becomes extremely difficult to avoid metastability.

REFERENCES

[1] S. E. Meninger and M. H. Perrott, "A 1-MHz Bandwidth 3.6-GHz 0.18-μm CMOS Fractional-N Synthesizer Utilizing a Hybrid PFD/DAC Structure for Reduced Broadband Phase Noise," *IEEE J. Solid-State Circuits,* vol. 41, pp. 966–981, April 2006.

[2] K. J. Wang, A. Swaminathan, and I. Galton, "Spurious Tone Suppression Techniques Applied to a Wide-Bandwidth 2.4 GHz Fractional-N PLL," *IEEE J. of Solid-State Circuits,* vol. 43, pp. 2787–2797, Dec. 2008.

[3] A. Zolfaghari, A. Y. Chan, and B. Razavi, "A 2.4-GHz 34-mW CMOS Transceiver for Frequency-Hopping and Direct-Sequence Applications," *ISSCC Dig. Tech. Papers,* pp. 418–419, Feb. 2001.

[4] G. Irvine et al., "An Upconversion Loop Transmitter IC for Digital Mobile Telephones," *ISSCC Dig. Tech. Papers,* pp. 364–365, Feb. 1998.

[5] T. Yamawaki et al., "A 2.7-V GSM RF Transceiver IC," *IEEE J. of Solid-State Circuits,* vol. 32, pp. 2089–2096, Dec. 1997.

[6] C. Lam and B. Razavi, "A 2.6-GHz/5.2-GHz Frequency Synthesizer in 0.4-μm CMOS Technology," *IEEE J. of Solid-State Circuits,* vol. 35, pp. 788–794, May 2000.

[7] C. S. Vaucher et al., "A Family of Low-Power Truly Modular Programmable Dividers in Standard 0.35-μm CMOS Technology," *IEEE J. of Solid-State Circuits,* vol. 35, pp. 1039–1045, July 2000.

[8] B. Razavi, Y. Ota, and R. G. Swartz, "Design Techniques for Low-Voltage High-Speed Digital Bipolar Circuits," *IEEE J. of Solid-State Circuits,* vol. 29, pp. 332–339, March 1994.

[9] J. Lee and B. Razavi, "A 40-Gb/s Clock and Data Recovery Circuit in 0.18-μm CMOS Technology," *IEEE J. of Solid-State Circuits,* vol. 38, pp. 2181–2190, Dec. 2003.

[10] D. D. Kim, K. Kim, and C. Cho, "A 94GHz Locking Hysteresis-Assisted and Tunable CML Static Divider in 65nm SOI CMOS," *ISSCC Dig. Tech. Papers,* pp. 460–461, Feb. 2008.

[11] J. Yuan and C. Svensson, "High-Speed CMOS Circuit Technique," *IEEE J. Solid-State Circuits,* vol. 24, pp. 62–70, Feb. 1989.

[12] B. Chang, J. Park, and W. Kim, "A 1.2-GHz CMOS Dual-Modulus Prescaler Using New Dynamic D-Type Flip-Flops," *IEEE J. Solid-State Circuits,* vol. 31, pp. 749–754, May 1996.

[13] R. L. Miller, "Fractional-Frequency Generators Utilizing Regenerative Modulation, " *Proc. IRE,* vol. 27, pp. 446–456, July 1939.

[14] J. Lee and B. Razavi, "A 40-GHz Frequency Divider in 0.18-μm CMOS Technology," *IEEE J. of Solid-State Circuits,* vol. 39, pp. 594–601, Apr. 2004.

[15] H. R. Rategh and T. H. Lee, "Superharmonic Injection-Locked Frequency Dividers," *IEEE J. of Solid-State Circuits,* vol. 34, pp. 813–821, June 1999.

[16] B. Razavi, "A Millimeter-Wave CMOS Heterodyne Receiver with On-Chip LO and Divider," *IEEE J. of Solid-State Circuits,* vol. 43, pp. 477–485, Feb. 2008.

[17] C.-C. Lin and C.-K. Wang, "A Regenerative Semi-Dynamic Frequency Divider for Mode-1 MB-OFDM UWB Hopping Carrier Generation," *ISSCC Dig. Tech. Papers,* pp. 206–207, Feb. 2005.

[18] B. Razavi, "A Study of Injection Locking and Pulling in Oscillators," *IEEE J. of Solid-State Circuits,* vol. 39, pp. 1415–1424, Sep. 2004.

PROBLEMS

10.1. Prove that insertion of a feedback divider in Fig. 10.14(b) results in Eqs. (10.25) and (10.26).

10.2. Prove that the far-out phase noise of the LO in Fig. 10.20 also appears as noise in the RX band. Neglecting other sources of noise, determine the phase noise at 25-MHz offset for GSM.

10.3. Does the sampling filter shown in Fig. 10.13(b) remove the effect of the mismatch between the Up and Down currents?

10.4. In practice, the sampling filter of Fig. 10.13(b) employs another capacitor tied from V_{cont} to ground. Explain why. How should this capacitor and C_2 be chosen to negligibly degrade the loop stability?

10.5. Suppose S_1 in Fig. 10.13(b) has a large on-resistance. Does this affect the loop stability?

10.6. Explain whether or not the charge injection and clock feedthrough of S_1 in Fig. 10.13(b) produce ripple on the control voltage after the loop has locked.

10.7. An in-loop modulation scheme is shown in Fig. 10.78. Consider two cases: the baseband bit period is (I) much shorter, or (II) much longer than the loop time constant.

(a) Sketch the output waveform of the VCO for both cases.

(b) Sketch the output waveform of the divider for both cases.

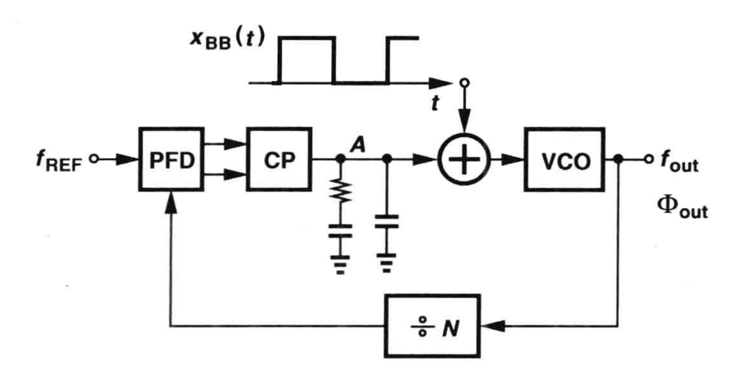


Figure 10.78 *PLL with in-loop modulation.*

10.8. In the pulse swallow divider of Fig. 10.24, the modulus control of the prescaler is accidentally inverted. Explain what happens.

10.9. In the PLL shown in Fig. 10.79, the control voltage has a sinusoidal ripple at a frequency of f_{REF}. Plot the spectra at A and B.

10.10. In the divider of Fig. 10.31, gate G_1 is mistakenly realized as a NAND gate. Explain what happens.

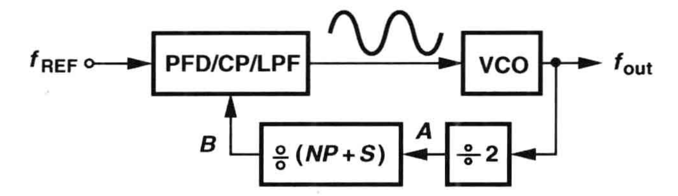


Figure 10.79 *PLL with divide-by-two circuit preceding the dual-modulus divider.*

10.11. In the circuit of Fig. 10.44(a), the resistors are replaced with ideal current sources. Explain what happens.

10.12. In our study of oscillators in Chapter 8, we concluded that a loop containing only two poles cannot oscillate (unless both are at the origin). Why then does the circuit of Fig. 10.49 oscillate?

10.13. Example 10.18 suggests that τ_{reg} is independent of R_D if $g_{m3,4}R_D \gg 1$. Explain this property intuitively.

10.14. Must the clock transition be abrupt for the D latch of Fig. 10.43(a) to operate properly? Consider a clock transition time (a) on the order of the time constant at X and Y, and (b) much longer than this time constant.

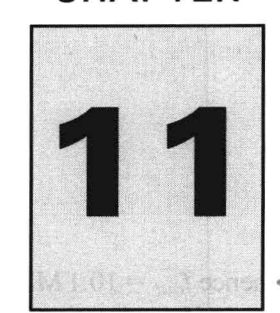
10.15. For the Miller divider of Fig. 10.68, determine the loop gain. Assume the tanks in the drains of M_1 and M_2 can be replaced by a resistance of R_p at resonance. Neglect all transistor capacitances and assume the resistors tied to the gates of M_5 and M_6 are large.

10.16. Prove that the Miller divider of Fig. 10.61(a) does not exhibit spurs at the output even if the mixer suffers from port-to-port feedthroughs.

10.17. Repeat the above problem if the mixer suffers from nonlinearity at each port.

10.18. Study the spurious response of the Miller divider shown in Fig. 10.69 if the mixer exhibits (a) port-to-port feedthrough, or (b) port nonlinearity.

10.19. Of the active mixer topologies studied in Chapter 6, which ones are suited to the Miller divider of Fig. 10.61(a)?



CHAPTER 11

FRACTIONAL-N SYNTHESIZERS

Our study of integer-N synthesizers in Chapter 10 points to a fundamental shortcoming of these architectures: the output channel spacing is equal to the reference frequency, limiting the loop bandwidth, settling speed, and the extent to which the VCO phase noise can be suppressed. "Fractional-N" architectures permit a *fractional* relation between the channel spacing and the reference frequency, relaxing the above limitations.

This chapter deals with the analysis and design of fractional-N synthesizers (FNS's). The chapter outline is shown below.

Randomization and Noise Shaping	Quantization Noise Reduction
Modulus Randomization	DAC Feedforward
Basic Noise Shaping	Fractional Divider
Higher-Order Noise Shaping	Reference Doubling
Out-of-Band Noise	Multiphase Division
Charge Pump Mismatch	

11.1 BASIC CONCEPTS

A PLL containing a $\div N$ circuit in the feedback multiplies the reference frequency by a factor of N. What happens if N is not *constant* with time? For example, what happens if the divider divides by N for half of the time and by $N + 1$ for the other half? We surmise that the "average" modulus of the divider is now equal to $[N + (N + 1)]/2 = N + 0.5$, i.e., the PLL, on the average, multiplies the reference frequency by a factor of $N + 0.5$. We also expect to obtain other fractional ratios between N and $N + 1$ by simply changing the percentage of the time during which the divider divides by N or $N + 1$.

As an example, consider the circuit shown in Fig. 11.1, where $f_{REF} = 1$ MHz and $N = 10$. Let us assume the prescaler divides by 10 for 90% of the time (nine reference cycles) and by 11 for 10% of the time (one reference cycle). Thus, for every 10 reference cycles, the output produces $9 \times 10 + 11 = 101$ pulses, yielding an average divide ratio of 10.1 and

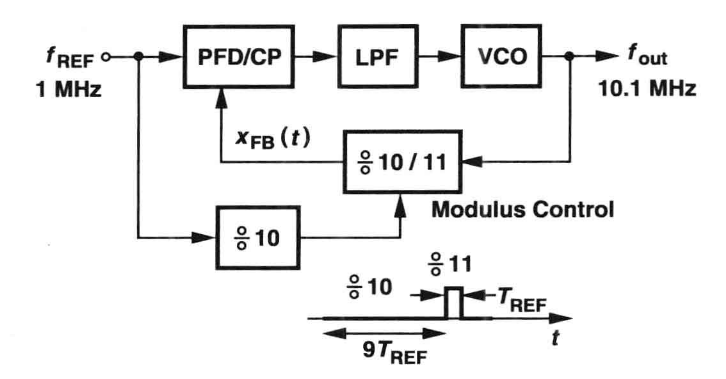


Figure 11.1 *Example of fractional-N loop.*

hence $f_{out} = 10.1$ MHz. In principle, the architecture can provide *arbitrarily* fine frequency steps if the durations of the $\div N$ and $\div(N+1)$ modes can be adjusted by small percentages.

The above example illustrates the efficacy of fractional-N synthesis with respect to creating fine channel spacings while running from a relatively high reference frequency. In addition to a wider loop bandwidth than that of integer-N architectures, this approach also reduces the in-band "amplification" of the reference phase noise (Chapter 10) because it requires a smaller N ($\approx f_{out}/f_{REF}$). (In the above example, an integer-N loop would multiply the reference phase noise by a factor of 100.)

The principal challenge in the design of FNS's stems from "fractional spurs." To understand this effect, let us return to the loop of Fig. 11.1 and reexamine its operation in the time domain. If the circuit operates as desired, then the output period is constant and equal to $(10.1\,\text{MHz})^{-1} \approx 99$ ns. Recall that, for nine reference cycles, this output period is multiplied by 10, and for one reference cycle, by 11. As shown in Fig. 11.2, each of the first nine cycles of the divided signal is 990 ns long, slightly *shorter* than the reference cycles. Consequently, the phase difference between the reference and the feedback signal grows in every period of f_{REF}, until it returns to zero when divide-by-11 occurs.[1] Thus, the phase detector generates progressively wider pulses, leading to a periodic waveform at the LPF output. Note that this waveform repeats every 10 reference cycles, modulating the VCO at a rate of 0.1 MHz and producing sidebands at $\pm 0.1\,\text{MHz} \times n$ around 10.1 MHz, where n denotes the harmonic number. These sidebands are called fractional spurs. More generally, for a nominal output frequency of $(N+\alpha)f_{REF}$, the LPF output exhibits a repetitive waveform with a period of $1/(\alpha f_{REF})$.

The appearance of fractional spurs can be explained from another perspective. Depicted in Fig. 11.3, the overall feedback signal, $x_{FB}(t)$ can be written as the sum of two waveforms, each of which repeats every 10,000 ns. The first waveform consists of nine periods of 990 ns and a "dead" time of 1090 ns, while the second is simply a pulse of width 1090/2 ns. Since each waveform repeats every 10,000 ns, its Fourier series consists of only harmonics at 0.1 MHz, 0.2 MHz, etc. If the phase detector is viewed as a mixer, we observe that the harmonics near 1 MHz are translated to "baseband" as they emerge from the PD, thus modulating the VCO.

1. This is only a simplistic view. Since a type-II PLL forces the *average* phase error to zero, the phase difference in fact fluctuates between positive and negative values.

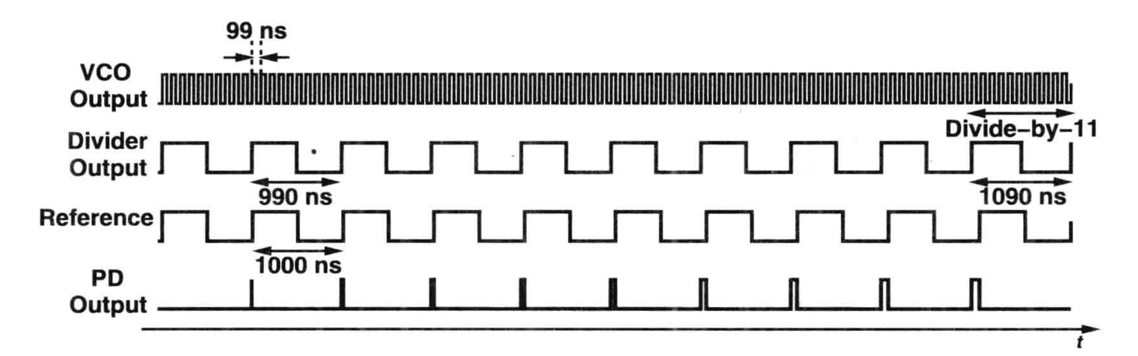


Figure 11.2 *Detailed operation of fractional-N loop.*

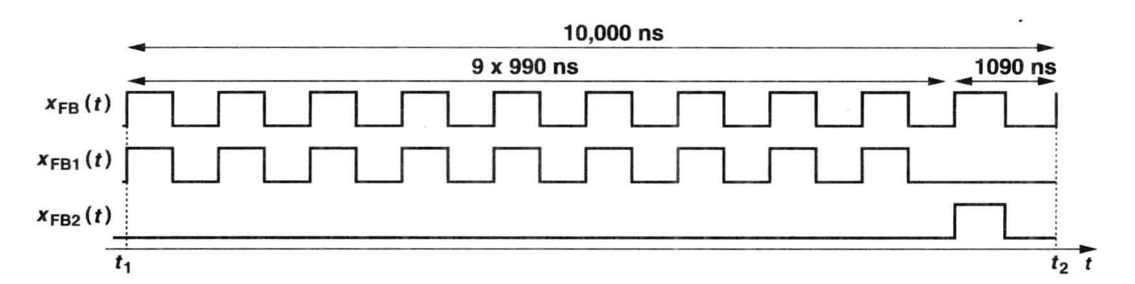


Figure 11.3 *Long periodicity in a fractional-N loop.*

Example 11.1

Determine the spectrum of $x_{FB1}(t)$ in Fig. 11.3.

Solution:

Let us first find the Fourier transform of one period of the waveform (from t_1 to t_2). This waveform consists of nine 990-ns cycles. If we had an infinite number of such cycles, the Fourier transform would contain only harmonics of 1.01 MHz. With nine cycles, the energy is spread out of the impulses, resembling that in Fig. 11.4(a). If this waveform is repeated every 10 μs, its Fourier transform is multiplied by a train of impulses located at integer multiples of 0.1 MHz. The spectrum thus appears as shown in Fig. 11.4(b).

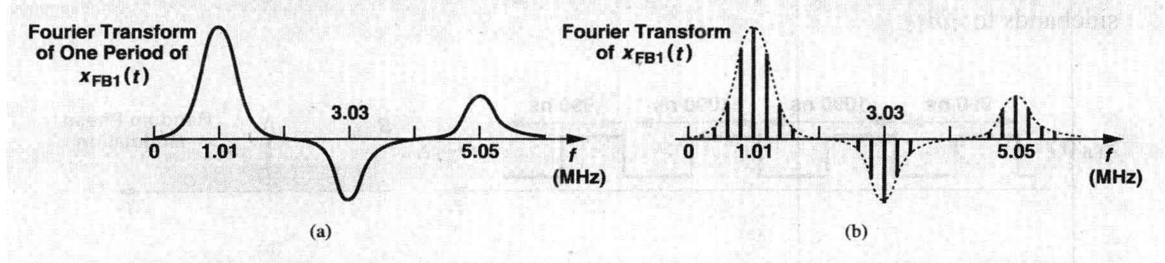


Figure 11.4 *(a) Fourier transform of one period of $x_{FB1}(t)$, (b) spectrum of $x_{FB1}(t)$.*

If the feedback signal is considered a 1-MHz waveform (while its fundamental frequency is in fact 0.1 MHz), then it contains many sidebands at integer multiples of 0.1 MHz. As explained in Chapter 3, the sidebands can be considered FM (and AM) components, leading to periodic phase modulation:

$$x_{FB}(t) \approx A\cos[\omega_{REF}t + \phi(t)]. \tag{11.1}$$

Comparing $x_{FB}(t)$ with an ideal reference at f_{REF}, the PFD produces an output proportional to $\phi(t)$, driving the loop filter with a periodic waveform at 0.1 MHz. This perspective does not need to consider the PFD as a mixer.

In summary, the feedback signal in the above example has a period of 10 μs and hence harmonics at $n \times 0.1$ MHz, but we roughly view it as a signal with an average frequency of 1 MHz and sidebands that are offset by ± 0.1 MHz, etc. These sidebands yield components at $n \times 0.1$ MHz at the PFD output, modulating the VCO and creating fractional spurs.

11.2 RANDOMIZATION AND NOISE SHAPING

The fractional spurs are quite large, requiring means of "compensation." The field of fractional-N synthesizers has introduced hundreds of compensation techniques in the past several decades. A class of techniques that lends itself to integration in CMOS technology and has become popular employs "noise shaping" [1]. This chapter is dedicated to this class of FNS's. We begin our study with the concepts of "modulus randomization" and noise shaping.

11.2.1 Modulus Randomization

Our analysis of the synthesizer in Fig. 11.1 reveals a periodicity in the behavior of the loop given by 10 reference cycles (0.1 MHz). What happens if the divider modulus is *randomly* set to 10 or 11 but such that its *average* value is still 10.1? As shown in Fig. 11.5(a), $x_{FB}(t)$ exhibits a random sequence of 990-ns and 1090-ns periods. Thus, unlike the situation portrayed in Fig. 11.4(b), $x_{FB}(t)$ now contains random phase modulation [Fig. 11.5(b)],

$$x_{FB}(t) = A\cos[\omega_{REF}t + \phi_n(t)], \tag{11.2}$$

leading to a random waveform (i.e., noise) at the PFD output. In other words, randomization of the modulus breaks the periodicity in the loop behavior, converting the deterministic sidebands to *noise*.

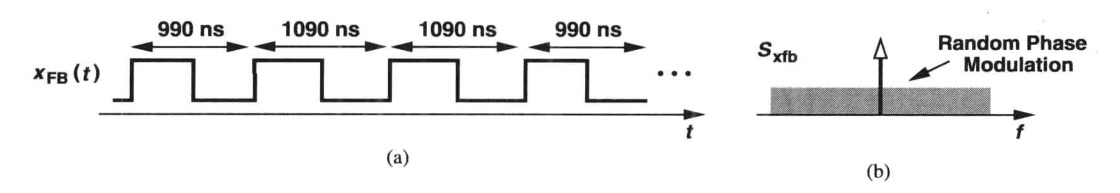


Figure 11.5 *(a) Randomization of divide ratio, (b) effect on spectrum of feedback signal.*

Let us now compute the noise, $\phi_n(t)$, in the feedback signal. Suppose the divider has two moduli, N and $N+1$, and must provide an average modulus of $N+\alpha$. We can write the instantaneous modulus as $N+b(t)$, where $b(t)$ randomly assumes a value of 0 or 1 and has an average value of α. The instantaneous frequency of the feedback signal is therefore expressed as

$$f_{FB}(t) = \frac{f_{out}}{N+b(t)}, \tag{11.3}$$

where f_{out} denotes the VCO output frequency. In the ideal case, $b(t)$ would be constant and equal to α, but our technique approximates α by a binary stream (i.e., with one-bit resolution), thereby introducing substantial noise. Since $b(t)$ is a random variable with a nonzero mean, we may write it in terms of its mean and another random variable with a zero mean:

$$b(t) = \alpha + q(t). \tag{11.4}$$

We call $q(t)$ the "quantization noise" because it denotes the error incurred by $b(t)$ in approximating the value of α. In Problem 11.1, we apply this result to the example in Fig. 11.1 with $N=10$, $\alpha=0.1$ but without randomization of $b(t)$.

Example 11.2

Plot $b(t)$ and $q(t)$ as a function of time.

Solution:

The sequence $b(t)$ contains an occasional square pulse so that the average is α [Fig. 11.6(a)]. Subtracting α from $b(t)$ yields the noise waveform, $q(t)$ [Fig. 11.6(b)].

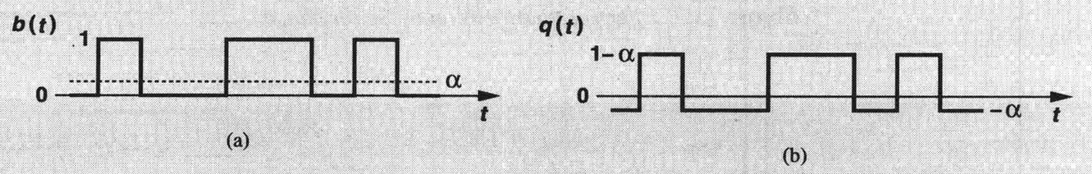


Figure 11.6 *(a) Random binary waveform having an average value of α, (b) quantization noise waveform.*

If $q(t) \ll N+\alpha$, we have

$$f_{FB}(t) = \frac{f_{out}}{N+\alpha+q(t)} \tag{11.5}$$

$$\approx \frac{f_{out}}{N+\alpha}\left[1-\frac{q(t)}{N+\alpha}\right] \tag{11.6}$$

$$\approx \frac{f_{out}}{N+\alpha} - \frac{f_{out}}{(N+\alpha)^2}q(t). \tag{11.7}$$

The feedback waveform arriving at the PFD is thus expressed as

$$V_{FB}(t) \approx V_0 \cos\left[\frac{2\pi f_{out}}{N+\alpha}t - \frac{2\pi f_{out}}{(N+\alpha)^2}\int q(t)dt\right], \tag{11.8}$$

because the phase is given by the time integral of the frequency. As expected, the divider output has an average frequency of $f_{out}/(N+\alpha)$ and a phase noise given by

$$\phi_{n,div}(t) = -\frac{2\pi f_{out}}{(N+\alpha)^2}\int q(t)dt. \tag{11.9}$$

In Problem 11.2, we compute this phase for the example in Fig. 11.1.

Example 11.3

Plot the phase noise in Eq. (11.9) as a function of time.

Solution:

With the aid of the waveform obtained in Example 11.2 for $q(t)$, we arrive at the random triangular waveform shown in Fig. 11.7.

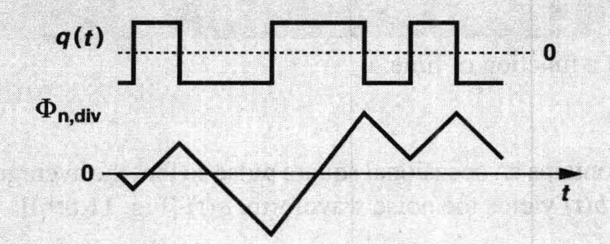


Figure 11.7 *Effect of quantization noise on phase.*

Example 11.4

Determine the spectrum of $\phi_{n,div}(t)$ from (11.9).

Solution:

The time integral of a function leads to a factor of $1/s$ in the frequency domain. Thus, the power spectral density of $q(t)$ must be multiplied by $[2\pi f_{out}/(N+\alpha)^2/\omega]^2$,

$$\overline{\phi^2_{n,div}}(f) = \frac{1}{(N+\alpha)^4}\left(\frac{f_{out}}{f}\right)^2 S_q(f), \tag{11.10}$$

where $S_q(f)$ is the spectrum of the quantization noise, $q(t)$. Note that this noise can be "referred" to the other PFD input—as if it existed in the reference waveform rather than the divider output.

Using the above results, we now determine the synthesizer output phase noise within the loop bandwidth. Viewing the phase noise as a component in the reference, we simply multiply Eq. (11.10) by the square of the average divide ratio, $N + \alpha$:

$$\overline{\phi_{n,out}^2} = \left[\frac{f_{out}}{(N+\alpha)f}\right]^2 S_q(f). \tag{11.11}$$

Alternatively, since $f_{out} = (N + \alpha)f_{REF}$,

$$\overline{\phi_{n,out}^2} = \left(\frac{f_{REF}}{f}\right)^2 S_q(f). \tag{11.12}$$

Example 11.5

Compute $S_q(f)$ if $b(t)$ consists of square pulses of width T_b that randomly repeat at a rate of $1/T_b$.

Solution:

We first determine the spectrum of $b(t)$, $S_b(f)$. As shown in Appendix I, $S_b(f)$ is given by

$$S_b(f) = \frac{\alpha(1-\alpha)}{T_b}\left(\frac{\sin \pi T_b f}{\pi f}\right)^2 + \alpha^2\delta(f), \tag{11.13}$$

where the second term signifies the dc content. Thus,

$$S_q(f) = \frac{\alpha(1-\alpha)}{T_b}\left(\frac{\sin \pi T_b f}{\pi f}\right)^2. \tag{11.14}$$

Figure 11.8 plots the spectrum, revealing a main "lobe" between $f = 0$ and $f = 1/T_b$. Note that, as T_b decreases, $S_q(f)$ contracts vertically and expands horizontally, maintaining a constant area under it (why?). In Problem 11.3, we consider this spectrum vis-a-vis the synthesizer loop bandwidth.

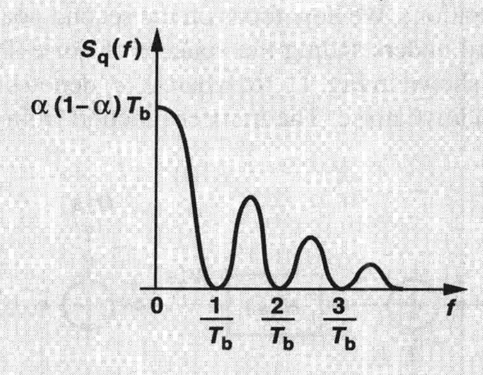


Figure 11.8 *Spectrum of random binary waveform.*

11.2.2 Basic Noise Shaping

While suppressing the fractional spurs, modulus randomization gives rise to a high phase noise. The high quantization noise arises from approximating a precise value, α, by only two coarse levels, namely, 0 and 1. Since the prescaler modulus cannot assume any other value between N and $N+1$, the resolution is limited to 1 bit, and the quantization noise cannot be reduced directly. Modern FNS's cope with this issue by performing the randomization such that the resulting phase noise exhibits a *high-pass* spectrum. Illustrated in Fig. 11.9, the idea is to minimize the spectral density near the center frequency of the feedback signal and allow the limited synthesizer loop bandwidth to suppress the noise farther away from the center frequency. The generation of the sequence $b(t)$ so as to create a high-pass phase spectrum is called "noise shaping."

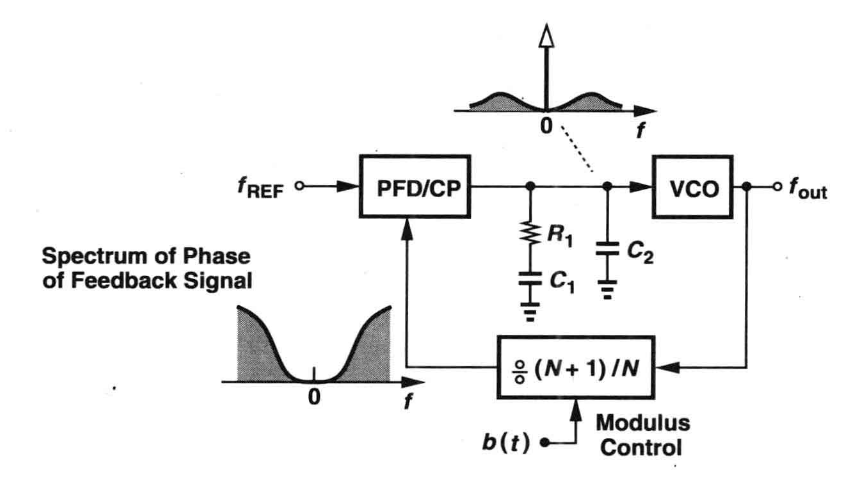


Figure 11.9 *Synthesizer employing modulus randomization.*

Let us summarize our thoughts. We wish to generate a random binary sequence, $b(t)$, that switches the divider modulus between N and $N+1$ such that (1) the average value of the sequence is α, and (2) the noise of the sequence exhibits a high-pass spectrum. The first goal is fulfilled if the number of ONEs divided by the number of ONEs and ZEROs is equal to α (over a long duration). We now focus on the second goal.

In our first step toward understanding the concept of noise shaping, we consider the negative feedback system shown in Fig. 11.10, where $X(s)$ denotes the main input and $Q(s)$ a secondary input, e.g., additive noise. The transfer function from $Q(s)$ to $Y(s)$ [with $X(s)$

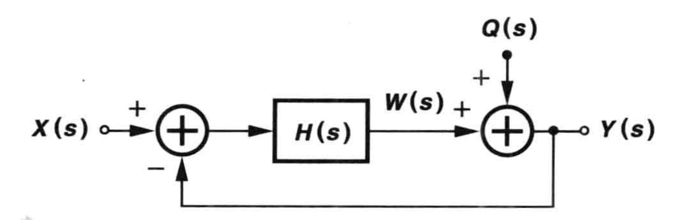


Figure 11.10 *Feedback system with noise injected near the output.*

set to zero] is equal to

$$\frac{Y(s)}{Q(s)} = \frac{1}{1 + H(s)}. \tag{11.15}$$

For example, if $H(s)$ is an ideal integrator,

$$\frac{Y(s)}{Q(s)} = \frac{s}{s + 1}. \tag{11.16}$$

In other words, a negative feedback loop containing an integrator acts as a *high-pass* system on the noise injected "near" the output. The reader may recognize the similarity of this behavior to the effect of VCO phase noise in PLLs (Chapter 9). If Q varies slowly with time, then the loop gain is large, making W a close replica of Q and hence Y small. From another point of view, the integrator provides a high loop gain at low frequencies, forcing Y to be approximately equal to X. Note that these results remain valid whether the system is analog, digital, or a mixture of analog and digital blocks.

Example 11.6

Construct a discrete-time version of the system shown in Fig. 11.10 if H must operate as an integrator.

Solution:

Discrete-time integration can be realized by *delaying* the signal and adding the result to itself [Fig. 11.11(a)]. We observe that if, for example, $A = 1$, then the output continues to rise in unity increments in each clock cycle. Since the z-transform of a single-clock delay

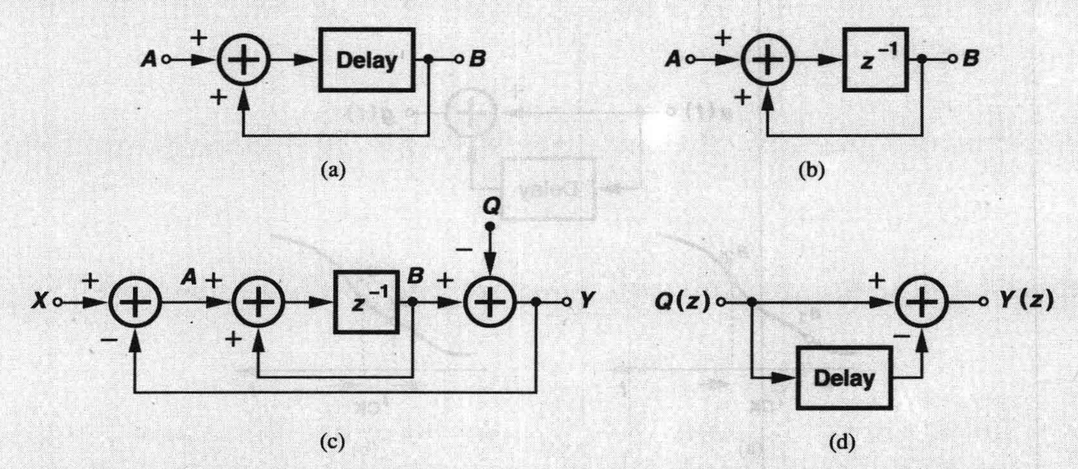


Figure 11.11 *(a) Discrete-time integrator, (b) integrator z-domain model, (c) use of integrator in a feedback loop, (d) simplified diagram of the feedback loop.*

(Continues)

Example 11.6 *(Continued)*

is equal to z^{-1}, we draw the integrator as shown in Fig. 11.11(b) and express the integrator transfer function as

$$\frac{B}{A}(z) = \frac{z^{-1}}{1 - z^{-1}}. \tag{11.17}$$

Thus, the discrete-time version of the system in Fig. 11.10 appears as shown in Fig. 11.11(c). Here, if $Q = 0$, then

$$\frac{Y}{X}(z) = z^{-1}, \tag{11.18}$$

i.e., the output simply tracks the input with a delay. Also, if $X = 0$, then

$$\frac{Y}{Q}(z) = 1 - z^{-1}. \tag{11.19}$$

This is a high-pass response (that of a differentiator) because, as conceptually illustrated in Fig. 11.11(d), subtracting the delayed version of a signal from the signal yields a small output if the signal does not change significantly during the delay.

The last point in the above example merits further investigation. Shown in Fig. 11.12 is a system that subtracts a delayed version of $a(t)$ from $a(t)$. The delay is equal to one clock cycle, T_{ck}. Figure 11.12(a) depicts a case where $a(t)$ changes significantly during one clock cycle, leading to an appreciable value for $a_2 - a_1$. That is, if $a(t)$ changes slowly, $g(t) \approx 0$. If the clock frequency increases [Fig. 11.12(b)], $a(t)$ finds less time to change, and a_1 and a_2 exhibit a small difference (i.e., they are strongly correlated). The key result here is that the systems in Figs. 11.11(c) and (d) reject Q by a *greater amount* if the delay element is clocked *faster*.

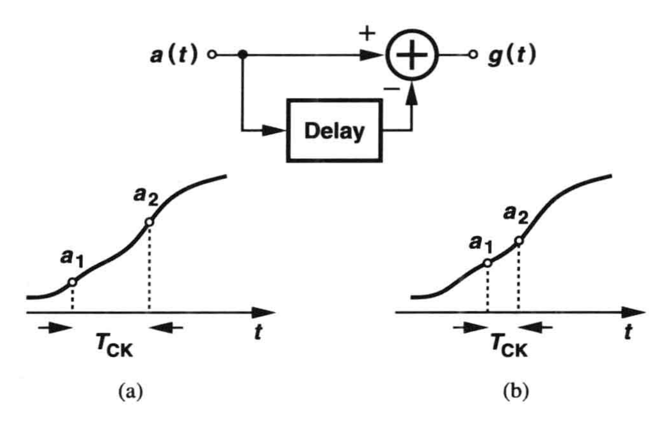


Figure 11.12 *Addition of a signal and its delayed version (a) for high and (b) low clock frequencies.*

Example 11.7

Construct the system of Fig. 11.11(c) in the digital domain with a precision (word length) of m bits.

Solution:

Shown in Fig. 11.13, the system incorporates an input adder (#1) (in fact a subtractor) and an integrator ("accumulator") consisting of a digital adder (#2) and a register (delay element). The first adder receives two m-bit inputs, producing an $(m + 1)$-bit output. Similarly, the integrator produces an $(m + 2)$-bit output. Since the feedback path from Y drops the two least significant bits of the integrator output, we say it introduces quantization noise, which is modeled by an additive term, Q.

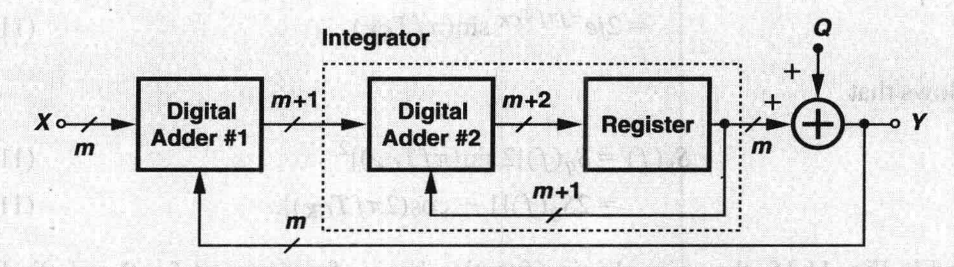


Figure 11.13 *Feedback system with an m-bit input.*

In analogy with the system shown in Fig. 11.10, we note that the high integrator gain forces Y to be equal to X at low frequencies, i.e., the *average* of Y is equal to the average of X.

We now assemble the concepts described thus far and construct a system that produces a binary output with an average value of α and a shaped noise spectrum. As shown in Fig. 11.14, we begin with an m-bit representation of α that is sufficiently accurate (X in Fig. 11.13). This value is applied to a feedback loop derived from that in Fig. 11.11(c), except that the high-resolution output of the integrator drives a flipflop (i.e., a one-bit quantizer), thereby generating a single-bit binary stream at the output. The quantization from

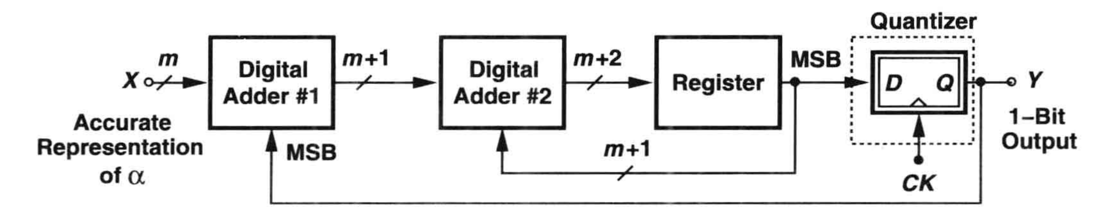


Figure 11.14 *ΣΔ modulator with one-bit output.*

$m + 2$ bits to 1 bit introduces significant noise, but the feedback loop shapes this noise in proportion to $1 - z^{-1}$. As explained in Example 11.7, the high integrator gain ensures that the average of the output is equal to X. This feedback system is called a "ΣΔ modulator." The choice of m in Fig. 11.14 is given by the accuracy with which the synthesizer output frequency must be defined. For example, for a frequency error of 10 ppm, $m \approx 17$ bits.

In the next step, we examine the shape of $1 - z^{-1}$ in the frequency domain. Recall from the definition of the z-transform that $z = \exp(j2\pi f T_{CK})$, where T_{CK} denotes the sampling or clock period. Thus, in the systems of Figs. 11.11(c) and 11.13,

$$\frac{Y}{Q}(z) = 1 - z^{-1} \tag{11.20}$$

$$= e^{-j\pi f T_{CK}} \left(e^{j\pi f T_{CK}} - e^{-j\pi f T_{CK}} \right) \tag{11.21}$$

$$= 2je^{-j\pi f T_{CK}} \sin(\pi f T_{CK}). \tag{11.22}$$

It follows that

$$S_y(f) = S_q(f)|2\sin(\pi f T_{CK})|^2 \tag{11.23}$$

$$= 2S_q(f)|1 - \cos(2\pi f T_{CK})|. \tag{11.24}$$

Plotted in Fig. 11.15, the noise shaping function begins from zero at $f = 0$ and climbs to 4 at $f = (2T_{CK})^{-1}$ (half the clock frequency). As predicted previously, a higher clock rate expands the function horizontally, thus reducing the noise density at low frequencies. The system of Fig. 11.14 is called a "first-order 1-bit ΣΔ modulator" because it contains one integrator.

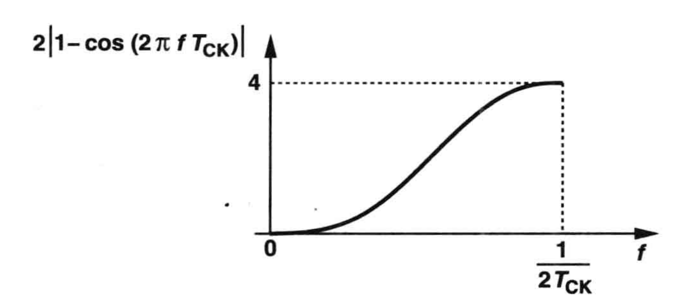


Figure 11.15 *Noise-shaping function in a first-order modulator.*

What can we say about the shape of $S_y(f)$? From Eq. (11.14),

$$S_y(f) = 2\frac{\alpha(1-\alpha)}{T_{CK}} \left(\frac{\sin \pi T_{CK} f}{\pi f} \right)^2 |1 - \cos(2\pi f T_{CK})|. \tag{11.25}$$

As explained below, the clock frequency, f_{CK}, is in fact equal to the synthesizer reference frequency, f_{REF}. Since the PLL bandwidth is much smaller than f_{REF}, we can consider $S_q(f)$ relatively flat for the frequency range of interest (Fig. 11.16). We hereafter assume that the shape of $S_y(f)$ is approximately the same as that of the noise-shaping function.

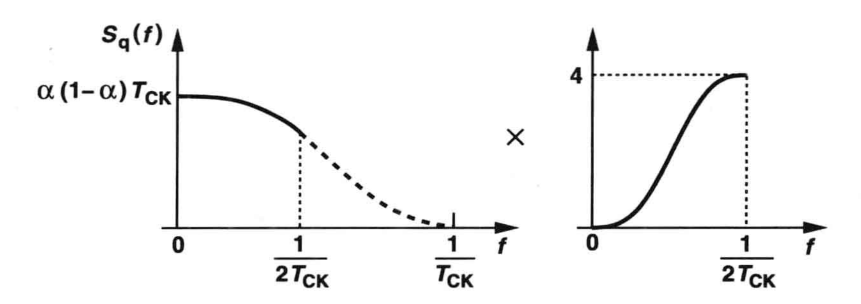


Figure 11.16 *Product of binary waveform quantization spectrum and noise-shaping function.*

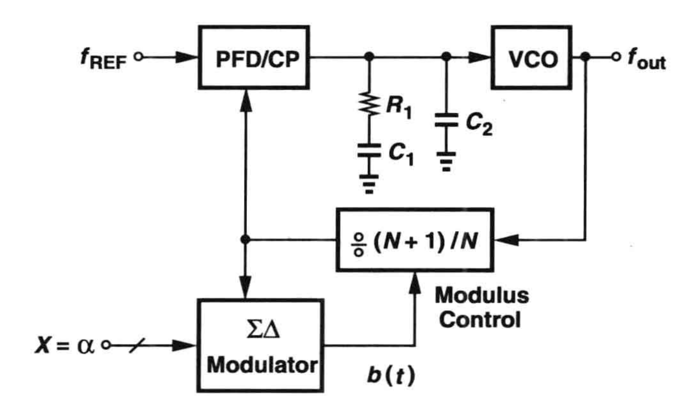


Figure 11.17 *Basic fractional-N loop using a ΣΔ modulator to randomize the divide ratio.*

Figure 11.17 shows the fractional-N synthesizer developed thus far. Clocked by the feedback signal, the ΣΔ modulator toggles the divide ratio between N and $N + 1$ so that the average is equal to $N + \alpha$.

Problem of Tones The output spectrum of ΣΔ modulators contains the shaped noise shown in Fig. 11.15, but also discrete *tones*. If lying at low frequencies, such tones are not removed by the PLL, thereby corrupting the synthesizer output.

To understand the origin of tones, we return to the modulator of Fig. 11.14 and ask, if X is constant, is the output binary sequence random? Since the system has no random inputs, we suspect that the output may not be random, either. For example, suppose $X = 0.1$. Then, as shown in Fig. 11.18, the output contains one pulse every ten clock cycles. In fact, after

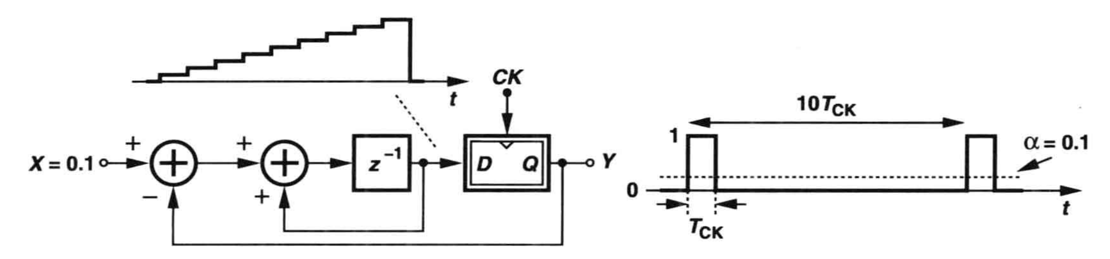


Figure 11.18 *Generation of idle tones in a ΣΔ loop.*

each output pulse, the integrator output falls to zero and subsequently rises in steps of 0.1 each clock cycle until it reaches 1 and drives the FF output high. In other words, the system exhibits a *periodic* behavior ("limit cycle"). Repeating every ten clock cycles, the output waveform of Fig. 11.18 consists of harmonics of $f_{CK}/10$, some of which are likely to fall within the bandwidth of the PLL and hence appear as spurs at the output.

To suppress these tones, the periodicity of the system must be broken. For example, if the LSB of X randomly toggles between 0 and 1, then the pulses in the output waveform of Fig. 11.18 occur randomly, yielding a spectrum with relatively small tones (but a higher noise floor). Called "dithering," this randomization must be performed at a certain rate: if excessively slow, it does not sufficiently break the periodicity. (Dithering may still produce tones in a nonlinear system.)

11.2.3 Higher-Order Noise Shaping

The noise shaping function expressed by Eq. (11.24) and illustrated in Fig. 11.15 does not adequately suppress the in-band noise. This can be seen by noting that, for $f \ll (\pi T_{CK})^{-1}$, Eq. (11.23) reduces to

$$S_y(f) = S_q(f)|2\pi f T_{CK}|^2; \tag{11.26}$$

i.e., the spectrum has a second-order roll-off as f approaches zero.[2] We therefore seek a system that exhibits a sharper roll-off, e.g., an output spectrum in proportion to f^n with $n > 2$. The following development will call for a "non-delaying integrator," shown in Fig. 11.19. The transfer function is given by

$$\frac{B}{A}(z) = \frac{1}{1 - z^{-1}}. \tag{11.27}$$

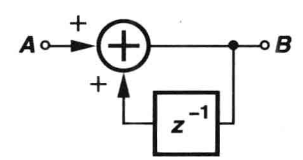


Figure 11.19 *Non-delaying integrator.*

In order to arrive at a system with a higher-order noise shaping, let us revisit the system of Fig. 11.14 and seek to increase the resolution of the quantizer (the flipflop) itself, i.e., a quantizer that produces lower quantization noise. From the foregoing developments, we recognize that a $\Sigma\Delta$ modulator can serve such a purpose because it suppresses the quantization noise at low frequencies. We therefore replace the 1-bit quantizer with a $\Sigma\Delta$ modulator [Fig. 11.20(a)]. To determine the noise shaping function, we write from the equivalent model shown in Fig. 11.20(b),

$$\left(-Y\frac{z^{-1}}{1 - z^{-1}} - Y\right)\frac{z^{-1}}{1 - z^{-1}} + Q = Y. \tag{11.28}$$

2. The in-band noise can also be reduced by raising f_{CK}, but in a synthesizer environment $f_{CK} = f_{REF}$.

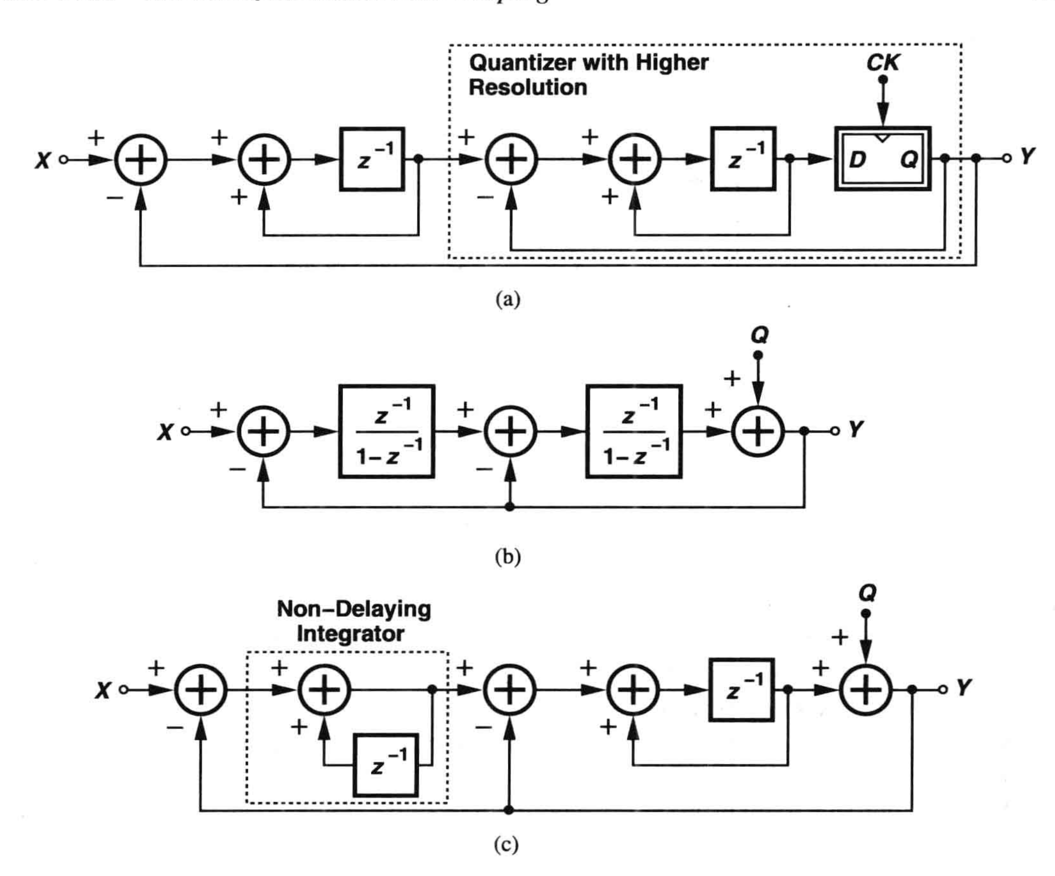


Figure 11.20 *(a) Use of a ΣΔ modulator as a quantizer within a ΣΔ loop, (b) simplified model of (a), (c) use of a non-delaying integrator.*

It follows that

$$\frac{Y}{Q}(z) = \frac{(1 - z^{-1})^2}{z^{-2} - z^{-1} + 1}. \tag{11.29}$$

The numerator indeed represents a sharper shaping, but the denominator exhibits two poles. Modifying the first integrator to a non-delaying topology [Fig. 11.20(c)], we have

$$\left(-Y\frac{1}{1 - z^{-1}} - Y\right)\frac{z^{-1}}{1 - z^{-1}} + Q = Y \tag{11.30}$$

and hence

$$\frac{Y}{Q}(z) = (1 - z^{-1})^2. \tag{11.31}$$

Following the derivations leading to Eq. (11.23), we have

$$S_y(f) = S_q(f)|2\sin(\pi f T_{CK})|^4, \tag{11.32}$$

i.e., the noise shaping falls in proportion to f^4 as f approaches zero. The system in Fig. 11.20(c) is called a "second-order 1-bit ΣΔ modulator." Plotted in Fig. 11.21 are the

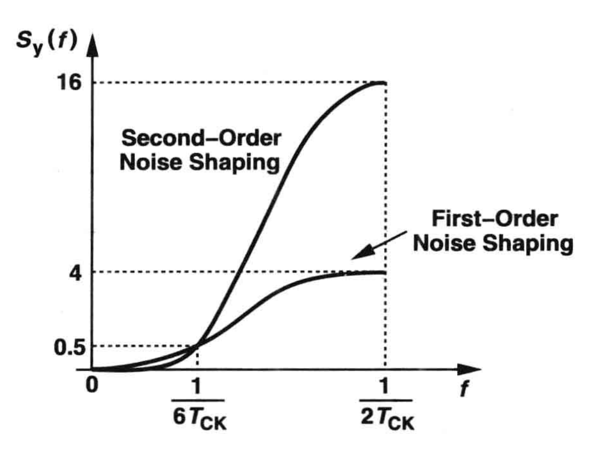


Figure 11.21 *Noise shaping in first- and second-order modulators.*

noise shaping functions given by (11.23) and (11.32), revealing that the latter remains lower than the former for frequencies up to $(6T_{CK})^{-1}$.

Is it possible to further raise the order of the $\Sigma\Delta$ modulator loop, thereby obtaining even sharper noise shaping functions? Yes, additional integrators in the loop provide a higher-order noise shaping. However, feedback loops containing more than two integrators are potentially unstable, requiring various stabilization techniques. Examples are described in [2].

Another approach to high-order $\Sigma\Delta$ modulator design employs "cascaded loops." Consider the first-order 1-bit loop shown in Fig. 11.22(a), where a subtractor finds the difference between the input and output of the quantizer, producing $U = Y_1 - W = Q$, i.e., the quantization error introduced by the quantizer. We postulate that if this error is *subtracted* from Y_1, the result contains a smaller amount of quantization noise. However, U has m bits. Thus, we must first convert it to a 1-bit representation with reasonable accuracy, a task well

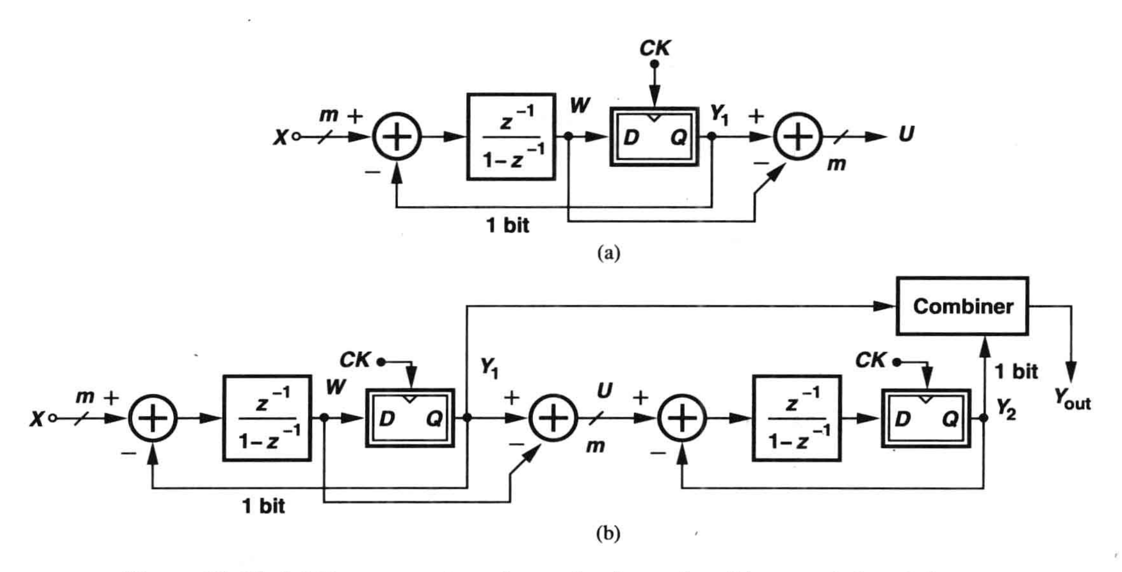


Figure 11.22 *(a) Reconstruction of quantization noise, (b) cascaded modulators.*

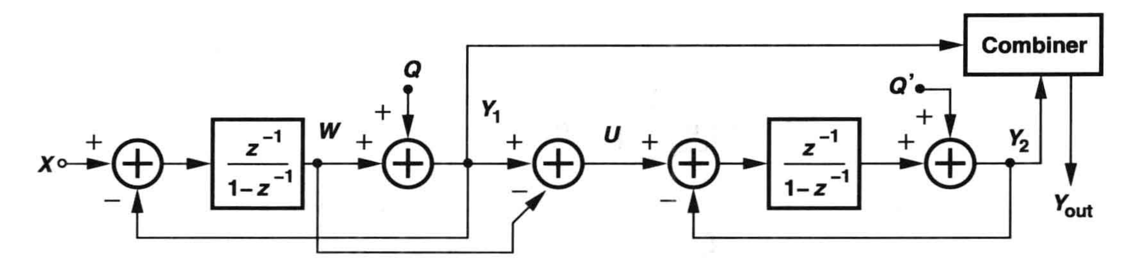


Figure 11.23 *Cascaded modulators showing quantization noise components.*

done by a ΣΔ modulator. As illustrated in Fig. 11.22(b), U drives a second loop, producing a 1-bit stream, Y_2. Since the quantization error due to approximating the m-bit U by the 1-bit Y_2 is shaped by the second loop, we observe that Y_2 is a relatively accurate replica of U. Lastly, Y_2 is combined with Y_1, yielding Y_{out} as a more accurate representation of X. This system is called a "1-1 cascade" to signify that each loop is of first order.

Let us compute the residual quantization noise present in Y_{out}. Redrawing the system of Fig. 11.22(b) as shown in Fig. 11.23, where Q' denotes the noise introduced by the second loop's quantizer, we have

$$Y_1(z) = z^{-1}X(z) + (1 - z^{-1})Q(z), \tag{11.33}$$

and

$$Y_2(z) = z^{-1}U(z) + (1 - z^{-1})Q'(z) \tag{11.34}$$

$$= z^{-1}Q(z) + (1 - z^{-1})Q'(z). \tag{11.35}$$

We wish to combine $Y_1(z)$ and $Y_2(z)$ such that $Q(z)$ is cancelled. To this end, the "combiner" multiplies both sides of (11.33) by z^{-1} and both sides of (11.35) by $(1 - z^{-1})$ and subtracts the latter result from the former:

$$Y_{out}(z) = z^{-1}Y_1(z) - (1 - z^{-1})Y_2(z) \tag{11.36}$$

$$= z^{-2}X(z) - (1 - z^{-1})^2 Q'(z). \tag{11.37}$$

Interestingly, the 1-1 cascade exhibits the same noise shaping behavior as the second-order modulator of Fig. 11.20(c).

Example 11.8

Construct a circuit that performs the combining operation shown in Fig. 11.23.

Solution:

For 1-bit streams, multiplication by z^{-1} is realized by a flipflop. The circuit thus appears as shown in Fig. 11.24.

(Continues)

Example 11.8 *(Continued)*

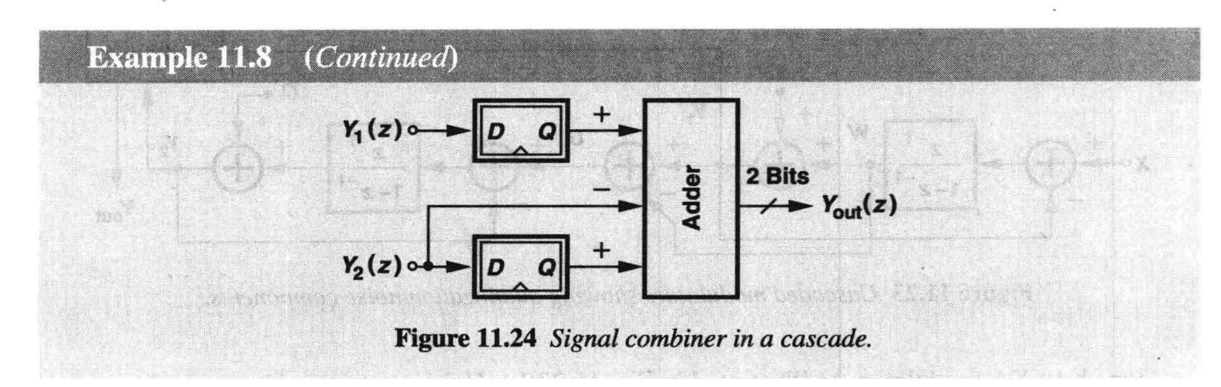


Figure 11.24 *Signal combiner in a cascade.*

Also known as the "MASH architecture," cascaded modulators can achieve high-order noise shaping without the risk of instability inherent in high-order single-loop modulators. However, as illustrated by the above example, the final output of a cascade is more than one bit wide, dictating a *multi*-modulus divider. For example, if Y_{out} assumes four possible levels, then a divider with moduli equal to $N-1$, N, $N+1$, and $N+2$ is necessary. Examples of multi-modulus dividers are described in [3, 4].

11.2.4 Problem of Out-of-Band Noise

The trend illustrated in Fig. 11.21 makes it desirable to raise the order of noise shaping so as to lower the in-band quantization noise. Unfortunately, however, higher orders inevitably lead to a sharper rise in the quantization noise at higher frequencies, a serious issue because the noise spectrum is multiplied by only a second-order low-pass transfer function as it travels to the PLL output.

To investigate this point, recall from Eq. (11.7) that the shaped noise spectrum expressed by Eq. (11.32), for example, is in fact *frequency* noise (as it represents modulation of the divide ratio). To compute the phase noise spectrum, we return to the transfer function from the quantization noise to the frequency noise:

$$Y(z) = (1 - z^{-1})^2 Q(z). \tag{11.38}$$

Now, since the phase noise, $\Phi(z)$, is the time integral of the frequency noise, $\Phi(z) = Y(z)/(1 - z^{-1})$,

$$\Phi(z) = (1 - z^{-1}) Q(z). \tag{11.39}$$

The spectrum of the phase noise is thus obtained as

$$S_\Phi(f) = |1 - z^{-1}|^2 S_q(f) \tag{11.40}$$

$$= |2\sin(\pi f T_{CK})|^2 S_q(f). \tag{11.41}$$

Appearing directly at one input of the phase detector, this phase noise spectrum is indistinguishable from the phase noise of the synthesizer reference, thus experiencing the low-pass transfer function of the PLL:

$$S_{out}(f) = |2\sin(\pi f T_{CK})|^2 S_q(f) N^2 \frac{4\zeta^2\omega_n^2\omega^2 + \omega_n^4}{(\omega^2 - \omega_n^2)^2 + 4\zeta^2\omega_n^2\omega^2}, \tag{11.42}$$

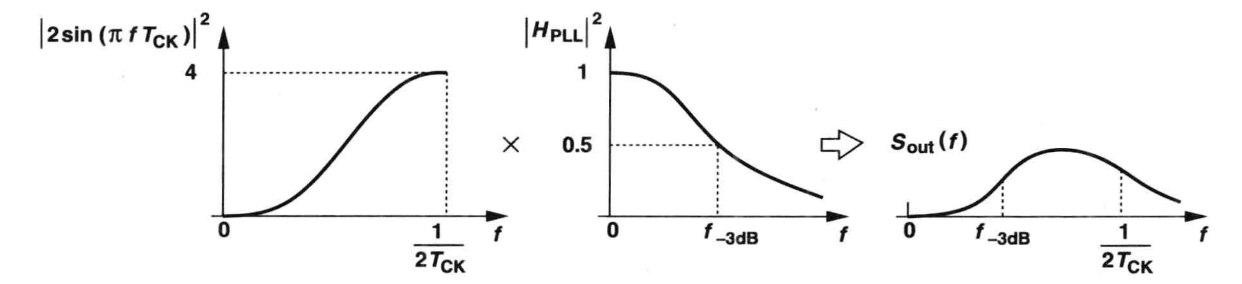


Figure 11.25 *Synthesizer output quantization noise.*

where N is the divider ratio and $\omega = 2\pi f$. Illustrated in Fig. 11.25 are the noise-shaped spectrum and the PLL transfer function.[3] If the $\Sigma\Delta$ modulator is clocked at a rate equal to the PLL reference (as is usually the case), then we note from Chapter 9 that $f_{-3dB} \approx 0.1 f_{REF} \approx 1/(10T_{CK})$. For small values of f, the noise shaping function in Eq. (11.42) can be approximated as $4\pi^2 f^2 T_{CK}^2$, whereas the PLL transfer function is equal to N^2. The product, $S_{out}(f)$, therefore begins from zero and rises to some extent. For larger values of f, the f^2 behavior of the noise shaping function cancels the roll-off of the PLL, leading to a relatively constant plateau. At values of f approaching $1/(2T_{CK}) = f_{REF}/2$, the product is dominated by the PLL roll-off. If comparable with the shaped VCO phase noise, this peaking of the $\Sigma\Delta$ phase noise spectrum proves troublesome. Figure 11.26 summarizes the phase noise effects at the synthesizer output.

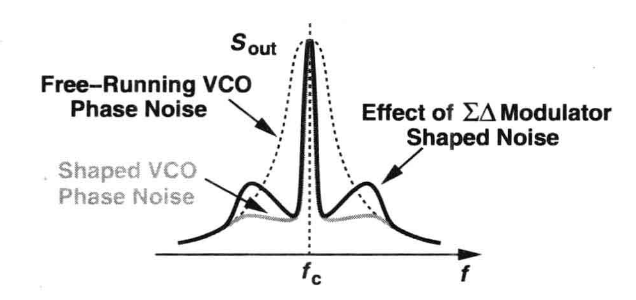


Figure 11.26 *Phase noise effects at the output of a fractional-N loop.*

The above study suggests that $\Sigma\Delta$ modulators having an order higher than 2 generate considerable phase noise at the synthesizer output unless the PLL bandwidth is reduced significantly, a trade-off violating the large bandwidth premise of fractional-N synthesizers. We quantify this behavior for a third-order modulator in Problem 11.4.

11.2.5 Effect of Charge Pump Mismatch

Our extensive study of PFD/CP nonidealities in Chapter 9 has revealed a multitude of effects that produce ripple on the control voltage of the oscillator and hence sidebands at the output. In particular, the mismatch between the Up and Down currents due to both

3. As explained in conjunction with Eq. (11.25), $S_q(f)$ is relatively flat in the frequency range of interest.

random effects and channel-length modulation proves quite serious in today's designs. This mismatch creates *additional* issues in fractional-N synthesizers [5].

In order to understand the effect of charge pump mismatch, we consider the PFD/CP/LPF combination shown in Fig. 11.27(a) and study the net *charge* delivered to C_1 as a function of the input phase difference, ΔT_{in} [5]. Note that the current sources are called I_1 and I_2 and the current waveforms arriving at the output node, I_{Up} and I_{Down}. Also, $I_{net} = I_{Up} - I_{Down}$. Depicted in Fig. 11.27(b) are the waveforms for the case where A leads B by ΔT_{in} seconds. The Up pulse goes high first, pumping a current of I_1. The Down pulse goes high ΔT_{in} seconds later, drawing a current of I_2, and lasts ΔT_1 seconds, where ΔT_1 denotes the PFD reset pulsewidth (about five gate delays). The net current, I_{net}, thus assumes a value of I_1 for ΔT_{in} seconds and a value of $I_1 - I_2$ for ΔT_1 seconds. Consequently, the total charge delivered to the loop filter is equal to

$$Q_{tot1} = I_1 \cdot \Delta T_{in} + (I_1 - I_2) \cdot \Delta T_1. \tag{11.43}$$

Now, let us reverse the polarity of the input phase difference. As shown in Fig. 11.27(c), the Down pulse goes high first, creating a net current of $-I_2$ until the Up pulse goes high and I_{net} jumps to $I_1 - I_2$. In this case,

$$Q_{tot2} = I_2 \cdot \Delta T_{in} + (I_1 - I_2)\Delta T_1. \tag{11.44}$$

(Note that ΔT_{in} is negative here.) The key observation here is that the slope of Q_{tot} as a function of ΔT_{in} jumps from I_2 to I_1 as ΔT_{in} goes from negative values to positive values

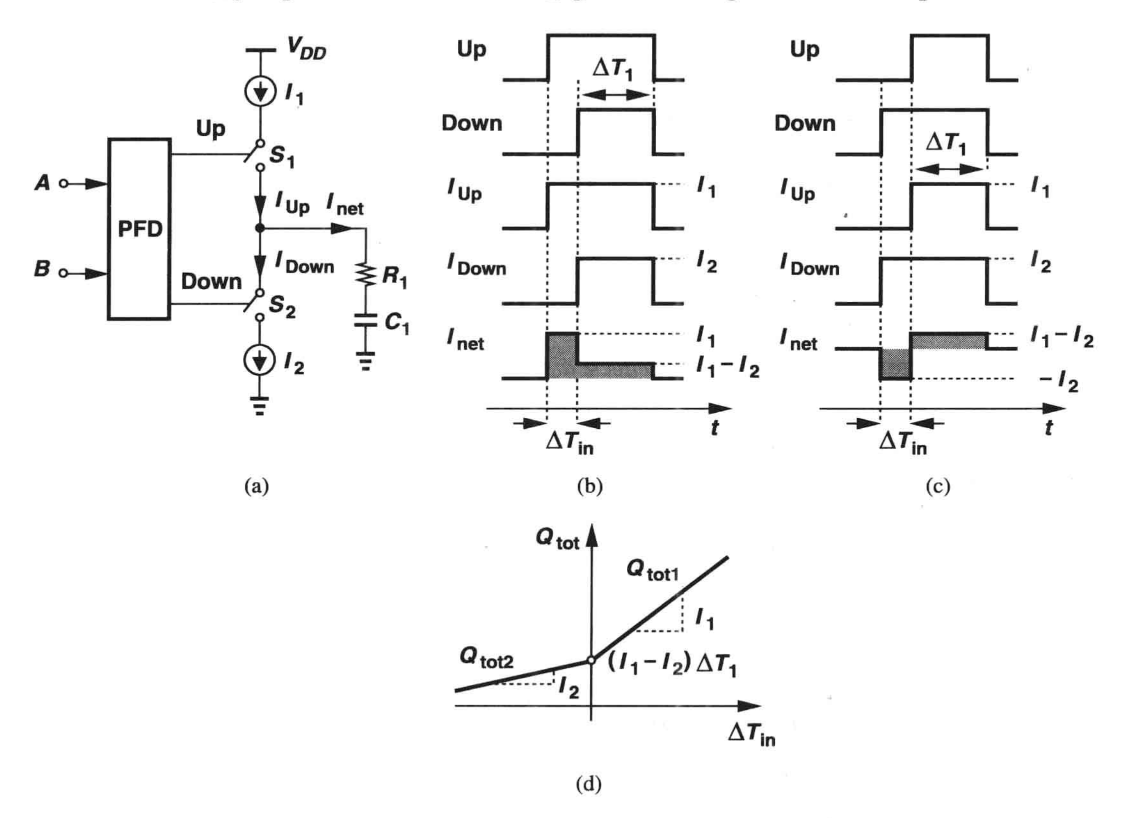


Figure 11.27 *(a) PFD/CP with current mismatches, (b) effect for Up ahead of Down, (c) effect for Up behind Down, (d) resulting characteristic.*

[Fig. 11.27(d)]. In other words, the PFD/CP characteristic suffers from *nonlinearity*. This nonlinearity adversely affects the broadband noise generated by the ΣΔ modulator and hence the feedback divider.

Example 11.9

Does the above nonlinearity manifest itself in integer-N synthesizers?

Solution:

No, it does not. Recall from Chapter 9 that, in the presence of a mismatch between I_1 and I_2, an integer-N PLL locks with a static phase offset, ΔT_0, such that the net charge injected into the loop filter is zero [Fig. 11.28(a)]. Now suppose the divider output phase

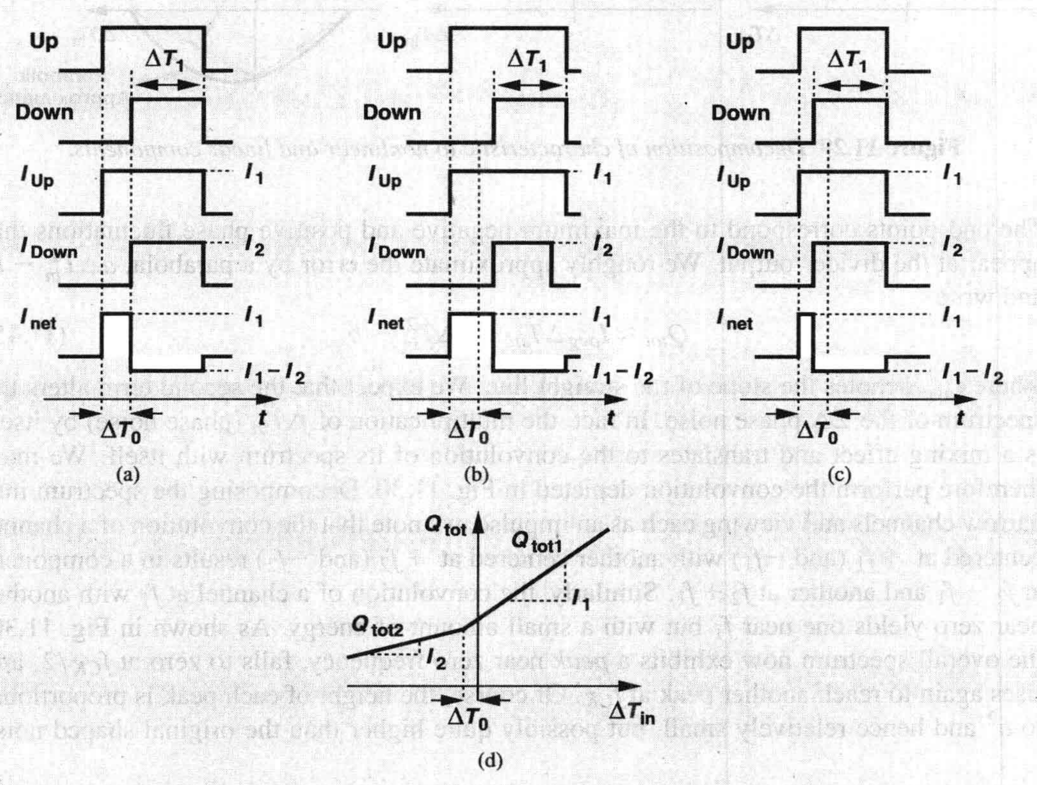


Figure 11.28 *Effect of current mismatch in an integer-N loop: (a) steady state, (b) random phase lead, (c) random phase lag, (d) resulting characteristic.*

experiences a small positive instantaneous jump (e.g., due to the VCO phase noise) [Fig. 11.28(b)]. The net charge therefore becomes proportionally positive. Similarly, for a small negative instantaneous phase jump, the net charge becomes proportionally negative [Fig. 11.28(c)]. The key point is that, in both cases, the charge is proportional to I_1, leading to the characteristic shown in Fig. 11.28(d). The nonlinearity is avoided so long as the jitter in the feedback signal remains less than ΔT_0. (If ΔT_0 is very small, so are the mismatch between I_1 and I_2 and hence the nonlinearity.)

The nonlinearity depicted in Fig. 11.27(d) becomes critical in $\Sigma\Delta$ fractional-N synthesizers because the feedback divider output contains *large*, random phase excursions. Since the phase difference sensed by the PFD fluctuates between large positive and negative values, the charge delivered to the loop filter is randomly proportional to I_1 or I_2.

What is the effect of the above nonlinearity on a $\Sigma\Delta$ fractional-N synthesizer? Let us decompose the characteristic of Fig. 11.27(d) into two components: a straight line passing through the "end points" and a nonmonotonic "error" (Fig. 11.29).

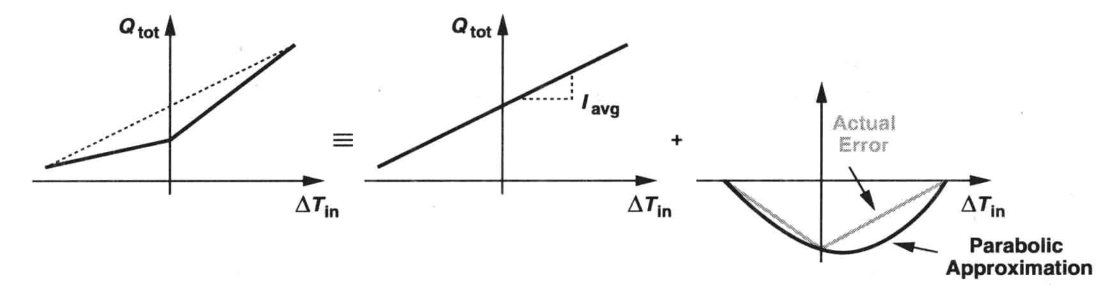


Figure 11.29 *Decomposition of characteristic to nonlinear and linear components.*

The end points correspond to the maximum negative and positive phase fluctuations that appear at the divider output. We roughly approximate the error by a parabola, $a\Delta T_{in}^2 - b$, and write

$$Q_{tot} \approx I_{avg}\Delta T_{in} + a\Delta T_{in}^2 - b, \tag{11.45}$$

where I_{avg} denotes the slope of the straight line. We expect that the second term alters the spectrum of the $\Sigma\Delta$ phase noise. In fact, the multiplication of ΔT_{in} (phase noise) by itself is a mixing effect and translates to the convolution of its spectrum with itself. We must therefore perform the convolution depicted in Fig. 11.30. Decomposing the spectrum into narrow channels and viewing each as an impulse, we note that the convolution of a channel centered at $+f_1$ (and $-f_1$) with another centered at $+f_2$ (and $-f_2$) results in a component at $f_2 - f_1$ and another at $f_2 + f_1$. Similarly, the convolution of a channel at f_1 with another near zero yields one near f_1 but with a small amount of energy. As shown in Fig. 11.30, the overall spectrum now exhibits a *peak* near zero frequency, falls to zero at $f_{CK}/2$, and rises again to reach another peak at f_{CK}. Of course, the height of each peak is proportional to a^2 and hence relatively small, but possibly quite higher than the original shaped noise

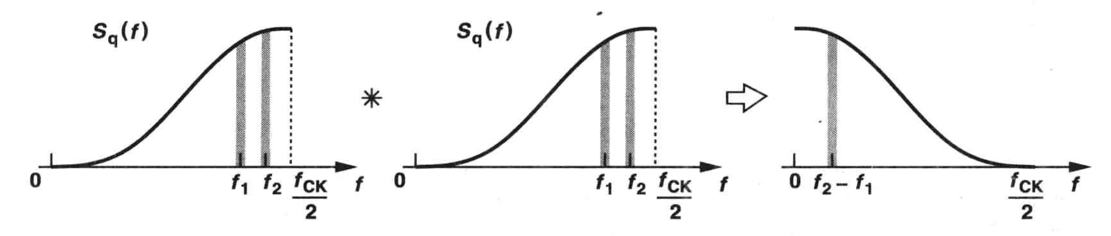


Figure 11.30 *Downconversion of high-frequency quantization noise as a result of CP nonlinearity.*

near zero frequency. That is, charge pump nonlinearity translates the ΣΔ modulator's high-frequency quantization noise to in-band noise, thus modulating the VCO. We also note that this "noise folding" effect becomes more pronounced as the order of the ΣΔ modulator and hence the high-frequency quantization noise increase. Similar folding may also occur for the idle tones described in Section 11.2.2.

In order to alleviate the charge pump mismatch issue, we can consider some of the solutions studied in Chapter 9. For example, the topology of Fig. 9.53 suppresses both random and deterministic mismatches, but it requires an op amp with a nearly rail-to-rail input common-mode range. Alternatively, we can return to Example 11.9 and observe that the nonlinearity does not manifest itself so long as the static phase offset is greater than the phase fluctuations produced by the feedback divider. In other words, if a *deliberate* mismatch is introduced between the Up and Down currents so as to establish a large static phase offset, then the slope of the characteristic remains constant around zero. As shown in Fig. 11.28(d), a mismatch of $\Delta I = I_1 - I_2$ affords a peak-to-peak phase fluctuation of

$$\Delta T_0 = \frac{\Delta I}{I} \Delta T_1, \tag{11.46}$$

where I denotes the smaller of I_1 and I_2 and ΔT_1, the width of the PFD reset pulses. Unfortunately, such a large mismatch also leads to a large ripple on the control voltage and a higher charge pump noise (because the CP current flows for a longer time).

Another approach to creating a static phase error splits the PFD reset pulse as depicted in Fig. 11.31(a) [6]. Since FF_1 is reset later than FF_2 by an amount equal to T_D, the Up pulse falls T_D seconds later than the Down pulse. The PLL must lock with a zero net charge, thus settling such that

$$\Delta T_0 \cdot I_2 \approx T_D \cdot I_1. \tag{11.47}$$

The static phase offset is now given by

$$\Delta T_0 \approx T_D. \tag{11.48}$$

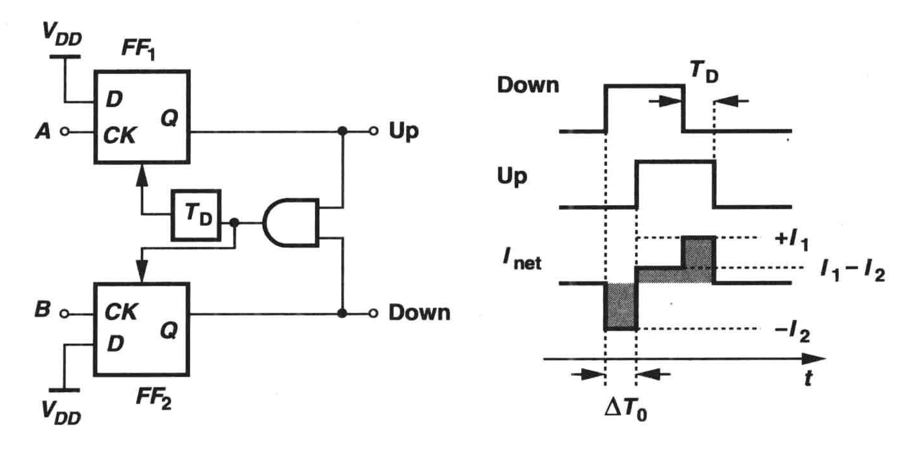


Figure 11.31 *Split reset pulses in a PFD to avoid slope change.*

That is, for a sufficiently large T_D and hence ΔT_0, phase fluctuations simply modulate the width of the negative current pulse in I_{net}, leading to a characteristic with a slope of I_2. Unfortunately, this technique also introduces significant ripple (a peak voltage of $I_2 \Delta T_0$) on the control voltage.

The above two approaches cope with the charge pump mismatch problem while producing ripple on the control voltage. Fortunately, as mentioned in Chapter 10, a sampling circuit interposed between the charge pump and the loop filter can "mask" the ripple, ensuring that the oscillator control line sees only the *settled* voltage produced by the CP (Fig. 11.32). In other words, a deliberate current offset or Up/Down misalignment along with a sampling circuit removes the nonlinearity resulting from the charge pump and yields a small ripple.

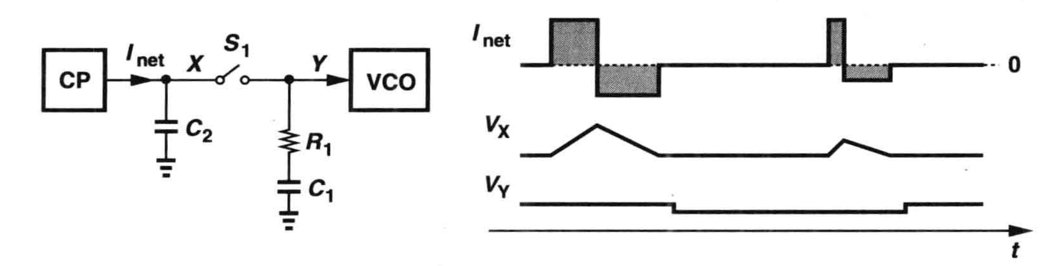


Figure 11.32 *Sampling filter to mask the control voltage from charge pump activity.*

11.3 QUANTIZATION NOISE REDUCTION TECHNIQUES

As explained in Section 11.2.4, the sharp rise in the quantization noise of $\Sigma\Delta$ modulators leads to substantial phase noise at the output of fractional-N synthesizers. In fact, this phase noise contribution can well exceed that of the VCO itself. This issue is ameliorated by reducing the PLL bandwidth, but at the cost of the advantages envisioned for fractional-N operation, namely, fast settling and suppression of the VCO phase noise across a large bandwidth. In this section, we study a number of techniques that lower the $\Sigma\Delta$ modulator phase noise contribution without reducing the synthesizer's loop bandwidth.

11.3.1 DAC Feedforward

Let us begin by *reconstructing* the quantization noise of the $\Sigma\Delta$ modulator. Figure 11.33 illustrates an example where a first-order, one-bit modulator produces

$$Y(z) = z^{-1}X(z) + (1 - z^{-1})Q(z). \tag{11.49}$$

We then delay $X(z)$ by one clock cycle and subtract the result from $Y(z)$ to reconstruct the quantization error:

$$W(z) = Y(z) - z^{-1}X(z) \tag{11.50}$$

$$= (1 - z^{-1})Q(z). \tag{11.51}$$

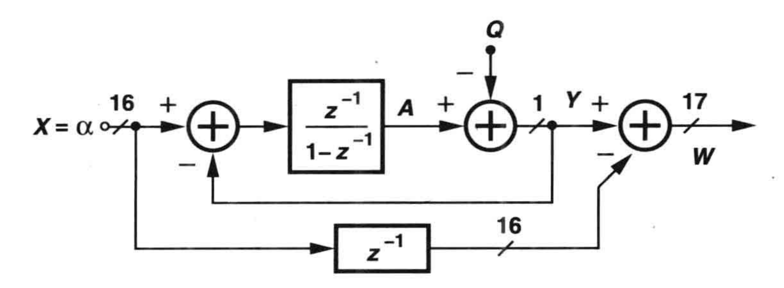


Figure 11.33 *Reconstruction of quantization noise.*

This operation yields the total (shaped) quantization error present in $Y(z)$. Note that, in this example, $W(z)$ has a 17-bit representation.

The reader must not confuse this operation with the quantization error reconstruction in cascaded modulators. Here, W is the shaped noise, whereas in cascaded modulators, we computed $Q = Y - A$, which is unshaped.

What can be done with the reconstructed error? Ideally, we wish to subtract $W(z)$ from $Y(z)$ to "clean up" the latter. However, such a subtraction would simply yield $X = \alpha$ with a 16-bit word length! We must therefore seek another point in the system where $W(z)$ can be subtracted from $Y(z)$, but not lead to a multi-bit digital signal.

Following the above line of thought, suppose, as shown in Fig. 11.34(a), we convert $W(z)$ to analog *charge* and inject the result into the loop filter with a polarity that cancels the effect of the $(1 - z^{-1})Q(z)$ noise arriving from the $\Sigma\Delta$ modulator. In the absence of analog and timing mismatches, each $\Sigma\Delta$ modulator output pulse traveling through the divider, the PFD, and the charge pump is met by another pulse produced by the DAC, facing perfect cancellation. We call this method "DAC feedforward cancellation."

The system of Fig. 11.34(a) entails a number of issues, requiring some modifications. First, we note that the PFD/CP combination generates a charge proportional to the *phase* of the divider output, i.e., the time integral of the frequency. Thus, the quantization noise arriving at the loop filter is of the form $(1 - z^{-1})Q(z)/(1 - z^{-1}) = Q(z)$, whereas the DAC output is of the form $(1 - z^{-1})Q(z)$. We must then interpose an integrator between the subtractor and the DAC. Figure 11.34(b) illustrates the result with a more general $\Sigma\Delta$ modulator of order L.

The second issue relates to the accuracy required of the DAC. Since it is extremely difficult to realize a 17-bit DAC, we may be tempted to truncate the DAC input to, say, 6 bits, but the truncation folds the high-frequency quantization noise to low frequencies [5] in a manner similar to the convolution shown in Fig. 11.30. It is thus necessary to "requantize" the 17-bit representation by another $\Sigma\Delta$ modulator, thereby generating a, say, 6-bit representation whose quantization noise is shaped [5] [Fig. 11.34(c)].

The third issue arises from the nature of the pulses travelling through the two paths. The Up and Down pulses activate the CP for only a fraction of the reference period, producing a current pulse of *constant* height each time. The DAC, on the other hand, generates current pulses of constant width. As shown in Fig. 11.35, the areas under the CP and DAC pulses are equal in the ideal case, but their arrival times and durations are not. Consequently, some ripple still appears on the control voltage. For this reason, the sampling loop filter of Fig. 11.32 is typically used to mask the ripple.

Figure 11.34 *(a) Basic DAC feedforward, (b) use of integrator in DAC path, (c) use of second modulator to relax required DAC resolution.*

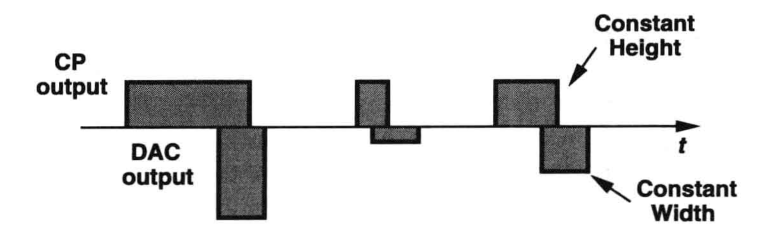


Figure 11.35 *CP and DAC output current waveforms.*

Example 11.10

What is the effect of the mismatch between the charge pump current and the DAC current in Fig. 11.34?

Solution:

The unequal areas of the current pulses generated by the CP and the DAC lead to incomplete cancellation of the quantization noise. For example, a 5% mismatch limits the noise reduction to roughly 26 dB.

The mismatch studied in the above example is also called the "DAC gain error." Figure 11.36(a) conceptually shows a 3-bit DAC whose output current is given by

$$I_{out} = I_{REF}(4D_3 + 2D_2 + D_1), \tag{11.52}$$

where $D_3D_2D_1$ represents the binary input. Figure 11.36(b) plots the input/output characteristic, revealing that an error in I_{REF} translates to an error in the *slope*, i.e., a gain error. Since both the charge pump current and the DAC current are defined by means of current mirrors, mismatches between these mirrors lead to incomplete cancellation of the quantization noise. Methods of DAC gain calibration are described in [7].

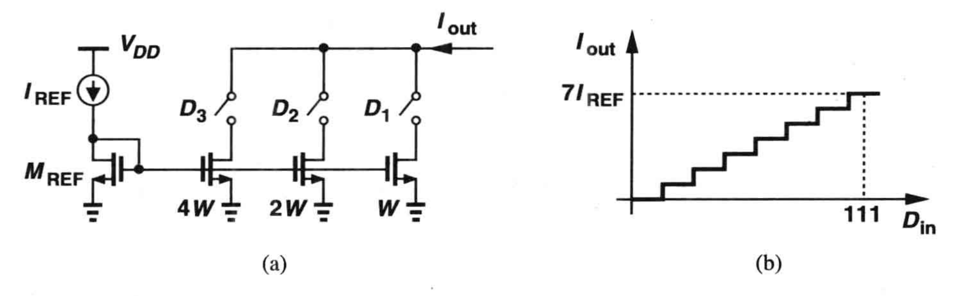


Figure 11.36 *(a) Current-mode DAC implementation, (b) input/output characteristic.*

How should the full-scale current of the DAC (e.g., $7I_{REF}$ in the above example) be chosen? The tallest pulse generated by the DAC must cancel the widest pulse produced by the CP, which in turn is given by the largest phase step at the output of the feedback divider. Interestingly, the maximum divider phase step depends on the order of the $\Sigma\Delta$ modulator, reaching three VCO cycles for an order of two and seven for an order of three. The DAC full scale is set accordingly.

In the feedforward approach described above, the DAC resolution need not exceed 5 or 6 bits, but its *linearity* must be quite higher [5]. Suppose, as shown in Fig. 11.37(a), the DAC characteristic exhibits some nonlinearity. Finding the difference between this characteristic and the straight line passing through the end points, we obtain the nonlinearity profile depicted in Fig. 11.37(b). In a manner similar to the study of charge pump nonlinearity in Section 11.2.5, we can approximate this profile by a polynomial, concluding that the DAC input is raised to powers of 2, 3, etc. As a result, the shaped high-frequency components of the quantization noise applied to the DAC are convolved and folded to low

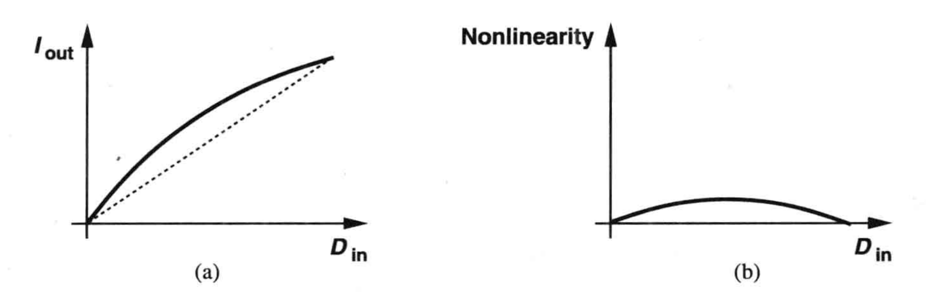


Figure 11.37 *(a) DAC characteristic and (b) nonlinearity profile.*

frequencies, raising the in-band phase noise (Fig. 11.30). For this reason, the DAC must employ additional measures to achieve a high linearity [5].

11.3.2 Fractional Divider

Another approach to reducing the $\Sigma\Delta$ modulator quantization noise employs "fractional" dividers, i.e., circuits that can divide the input frequency by noninteger values such as 1.5 or 2.5. For example, if a circuit can divide by 2 or 2.5, then the quantization error is halved, exhibiting a spectrum that is shifted down by 6 dB.[4] But how can a circuit divide by, say, 1.5? The key to this operation is the notion of "double-edge-triggered" (DET) flipflops. Illustrated in Fig. 11.38(a), a DET flipflop incorporates two D latches driven by CK and $\overline{CK}$ and a multiplexer (MUX). When CK is high, the top latch is in the sense mode and the bottom latch in the hold mode, and vice versa. Also, the MUX selects A when CK is low and B when it is high.[5] Let us now drive the circuit with a "half-rate clock," i.e., one whose period is twice the input bit period. Thus, as depicted in Fig. 11.38(b), even with a half-rate clock, D_{out} tracks D_{in}. In other words, for a given clock rate, the input data to a DET flipflop can be *twice* as fast as that applied to a single-edge-triggered counterpart. Figure 11.38(c) shows a CML realization of the circuit.

Let us now return to the ÷3 circuit studied in Chapter 9 and replace the flipflops with the DET circuit of Fig. 11.39(a).[6] Noting that each FF now "reads" its input both when CK is high and when it is low, we begin with $Q_1\overline{Q_2} = 00$ and observe that the first high clock level maintains Q_1 at ZERO (because $\overline{Q_2}$ was ZERO) and raises $\overline{Q_2}$ to ONE (because Q_1 was ZERO) [Fig. 11.39(b)]. When CK falls, the flipflops read their inputs again, producing $Q_1 = 1$ and $\overline{Q_2} = 1$. Finally, when CK goes high again, Q_1 remains high while $\overline{Q_2}$ falls. The circuit therefore produces one output period for every 1.5 input periods.

DET flipflops can be used in other dividers having an odd modulus to obtain a fractional divide ratio. For example, a ÷5 circuit is readily transformed to a ÷2.5 stage. Note, however, that DET flipflops suffer from a larger clock input capacitance than their single-edge-triggered counterparts. Also, clock duty cycle distortion leads to unwanted spurs at the output.

---

4. Since the *magnitude* of the error is halved, the PSD drops by 6 dB rather than 3 dB.

5. This choice of clock phases for the latches and the MUX allows master-slave operation between each latch and the MUX.

6. The double-edge operation is denoted by two hats at the clock input.

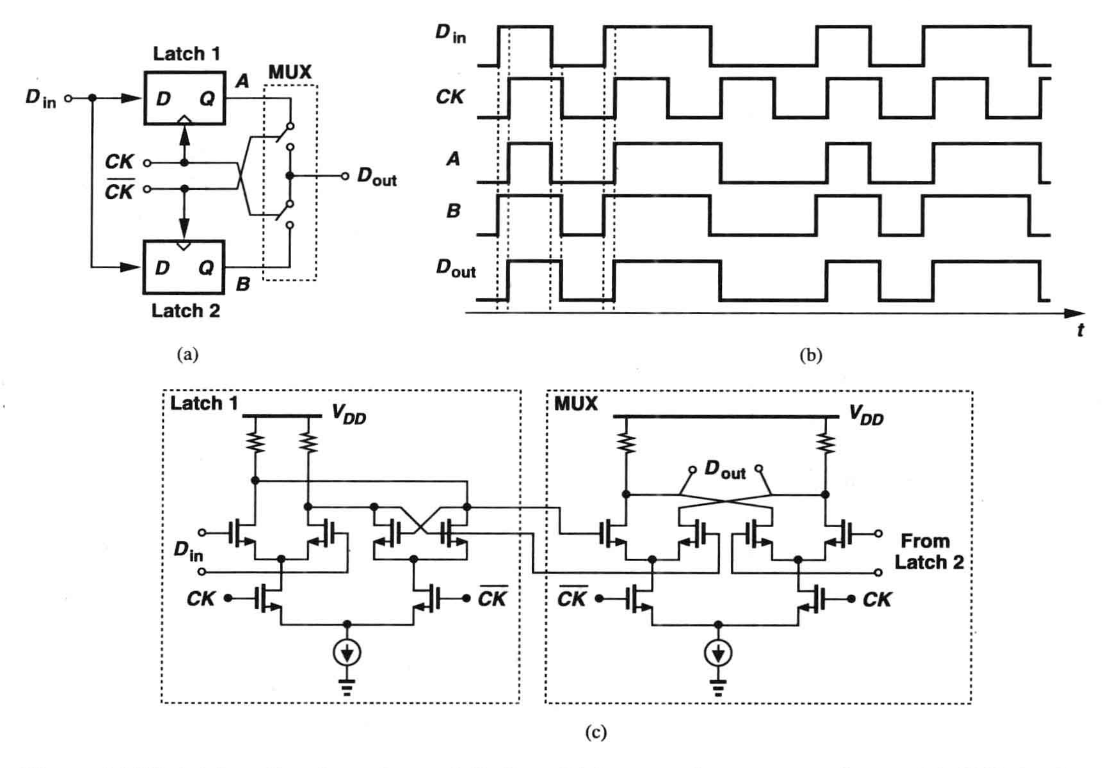


Figure 11.38 *(a) Double-edge-triggered flipflop, (b) input and output waveforms, (c) CML implementation.*

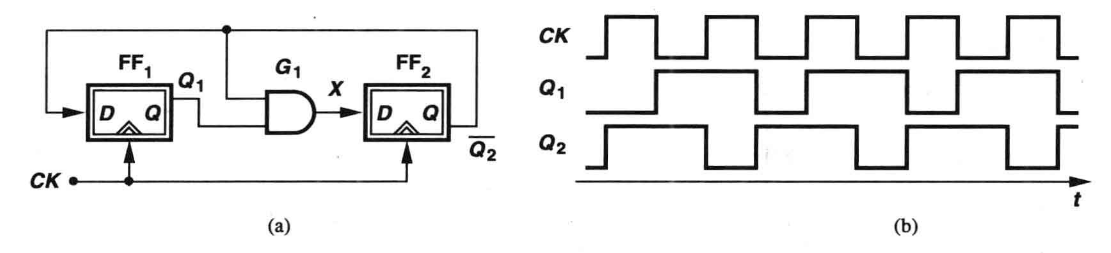


Figure 11.39 *(a) Divide-by-1.5 circuit, (b) input and output waveforms.*

11.3.3 Reference Doubling

Our derivation of the noise shaping function in Section 11.2.2 indicates a direct dependence on the clock frequency. In fact, Eq. (11.26) suggests that if T_{CK} is halved, the noise power falls by 6 dB, making it desirable to use the highest available reference frequency. Generated by a crystal oscillator, the reference frequency is typically limited to less than 100 MHz, especially if power consumption, phase noise, and cost are critical. We then surmise that if the reference frequency can be doubled by means of an on-chip circuit preceding the PLL, then the phase noise due to the $\Sigma\Delta$ modulator quantization can be reduced by 6 dB (for a first-order loop) [8].

Figure 11.40 shows a frequency doubler circuit: the input is delayed and XORed with itself, producing an output pulse each time $V_{in}(t)$ and $V_{in}(t - \Delta T)$ are unequal.

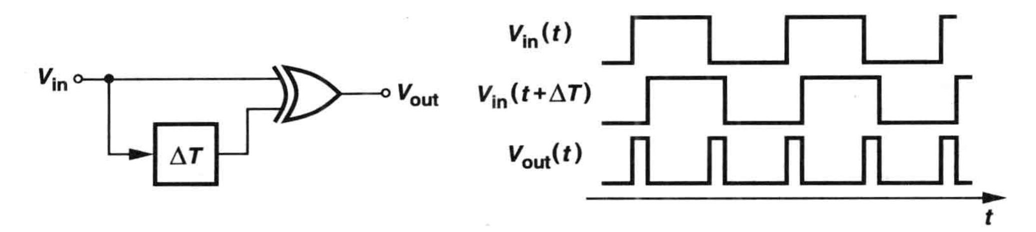


Figure 11.40 *Frequency doubler.*

Example 11.11

If we consider $V_{out}(t)$ in Fig. 11.40 as the sum of two half-rate waveforms (Fig. 11.41), determine the Fourier series of $V_{out}(t)$.

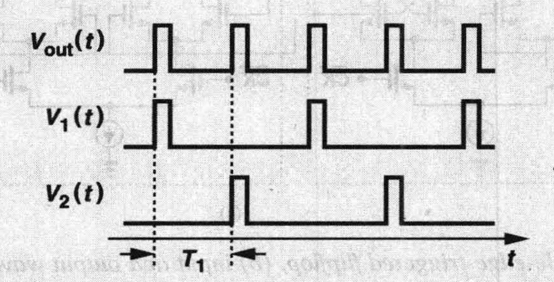


Figure 11.41 *Decomposition of doubler output.*

Solution:

The Fourier series of $V_1(t)$ can be written as

$$V_1(t) = a_1 \cos(\omega_0 t + \phi_1) + a_2 \cos(2\omega_0 t + \phi_2) + a_3 \cos(3\omega_0 t + \phi_3) + \cdots, \qquad (11.53)$$

where $\omega_0 = 2\pi/(2T_1)$. The second waveform, $V_2(t)$, is obtained by shifting V_1 by T_1. Thus, the first harmonic is shifted by $\omega_0 T_1 = \pi$, the second by $2\omega_0 T_1 = 2\pi$, etc. It follows that

$$V_2(t) = -a_1 \cos(\omega_0 t + \phi_1) + a_2 \cos(2\omega_0 t + \phi_2) - a_3 \cos(3\omega_0 t + \phi_3) + \cdots. \qquad (11.54)$$

Adding $V_1(t)$ and $V_2(t)$, we note that all odd harmonics of ω_0 vanish, yielding a waveform with a fundamental frequency of $2\omega_0$.

Unfortunately, the circuit of Fig. 11.40 produces unevenly-spaced pulses if the input duty cycle deviates from 50% [Fig. 11.42(a)]. Following the above example [8], we decompose the output into two waveforms having a period of $2T_1$ and recognize that the time shift between $V_1(t)$ and $V_2(t)$, ΔT, now deviates from T_1. Thus, the odd harmonics are not completely cancelled, appearing as sidebands around the main component at $1/T_1$ [Fig. 11.42(b)]. Since the PLL bandwidth is chosen about one-tenth of $1/T_1$, the sidebands are attenuated to some extent. We prove in Problem 11.10 that a loop bandwidth of

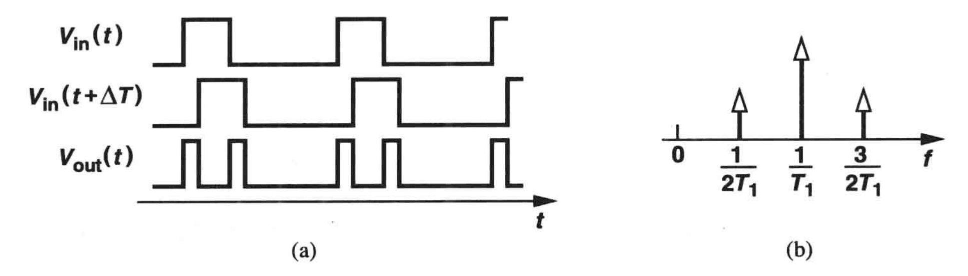


Figure 11.42 *(a) Doubler output in the presence of input duty cycle distortion, (b) resulting spectrum.*

$2.5\omega_n = 0.1 \times (2\pi/T_1)$ lowers the magnitude of the sidebands at $1/(2T_1)$ by about 16 dB. However, while traveling to the synthesizer output, the sidebands grow by a factor equal to the feedback divide ratio.

The foregoing analysis reveals that the duty cycle of the input waveform must be tightly controlled. The synthesizer described in [8] employs a duty cycle correction circuit. Such circuits still suffer from residual duty cycle errors due to their internal mismatches, possibly yielding unacceptably large reference sidebands at the synthesizer output—unless the loop bandwidth is reduced.

11.3.4 Multiphase Frequency Division

It is possible to reduce the quantization error in the divide ratio through the use of multiple phases of the VCO. From our analysis in Section 11.1, we note that when the divider modulus switches from N to $N+1$ (or vice versa), the divider output phase jumps by one VCO period (Fig. 11.43). On the other hand, if finer phases of the VCO are available, the phase jumps can become proportionally smaller, resulting in lower quantization noise.

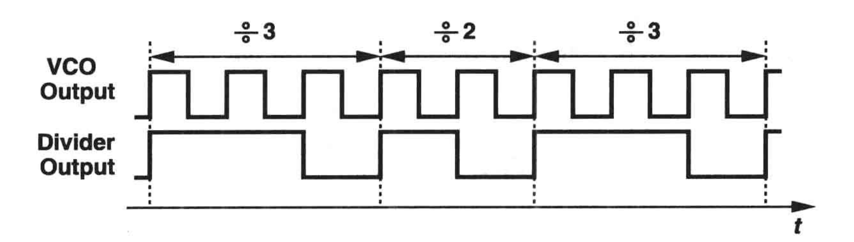


Figure 11.43 *Phase jumps at the output of dual-modulus divider.*

It is possible to create a fractional divide ratio by means of a multiphase VCO and a multiplexer. Suppose a VCO generates M output phases with a minimum spacing of $2\pi/M$, and the MUX selects one phase each time, producing an output given by

$$V_{MUX}(t) = V_0 \cos\left(\omega_c t - k\frac{2\pi}{M}\right), \tag{11.55}$$

where k is an integer. Now, let us assume that k varies linearly with time, sequencing through $0, 1, \ldots, M-1, M, M+1, \ldots$. Thus, $k = \beta t$, where β denotes the rate of change

of k, and hence

$$V_{MUX}(t) = V_0 \cos\left[\left(\omega_c - \beta\frac{2\pi}{M}\right)t\right], \tag{11.56}$$

revealing a frequency of $\omega_c - \beta(2\pi/M)$. The divide ratio is therefore equal to $1 - (\beta/\omega_c)(2\pi/M)$.

As an example, consider the circuit shown in Fig. 11.44(a), where the quadrature phases of a VCO are multiplexed to generate an output. Initially, V_I is selected and V_{out} tracks V_I until $t = t_1$, at which point, V_Q is selected. Similarly, V_{out} tracks V_Q until $t = t_2$, and then $\overline{V_I}$ until $t = t_3$, etc. We therefore observe that this periodic "stitching" of the quadrature phases yields an output with a period of $T_{in} + T_{in}/4 = 5T_{in}/4$, equivalently, a $\div 1.25$ operation. In other words, this technique affords a frequency divider having a modulus of 1 and a modulus of 1.25 [10]. Since the divide ratio can be adjusted in a step of 0.25, the quantization noise falls by $20 \log 4 = 12$ dB [10].

The use of *quadrature* LO phases in the above example does not pose additional constraints on the system because direct-conversion transceivers require such phases for upconversion and downconversion anyway. However, finer fractional increments necessitate additional LO phases, making the oscillator design more complex and power-hungry.

Multiphase fractional division must deal with two issues. First, the MUX select command (which determines the phase added to the carrier each time) is difficult to generate. This is because, to avoid glitches at the MUX output, this command must change only when none of the MUX inputs is changing. Viewing the MUX in Fig. 11.44(a) as four differential pairs whose output nodes are shared and whose tail currents are sequentially enabled, we draw the input and select waveforms of the $\div 1.25$ circuit as shown in Fig. 11.45. Note that the edges of the select waveforms have a small margin with respect to the input edges. Moreover, if the divide ratio must switch from 1.25 to 1, a different set of select waveforms must be applied, complicating the generation and routing of the select logic.

The second issue in multiphase fractional dividers relates to phase mismatches. In the circuit of Fig. 11.44(a), for example, the quadrature LO phases and the paths within the MUX suffer from mismatches, thereby displacing the output transitions from their ideal

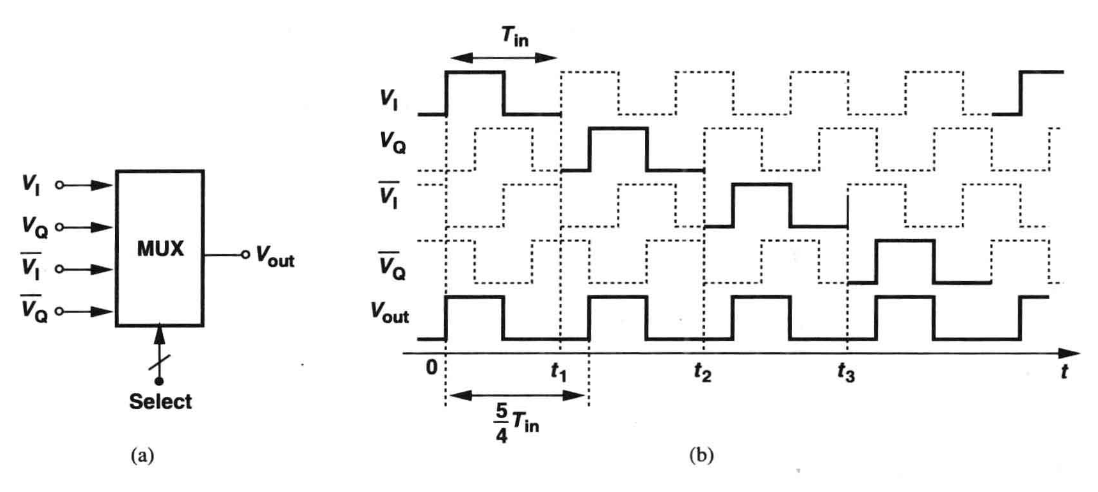


Figure 11.44 *(a) Multiplexed VCO phases, (b) waveforms showing divide-by-1.25 operation.*

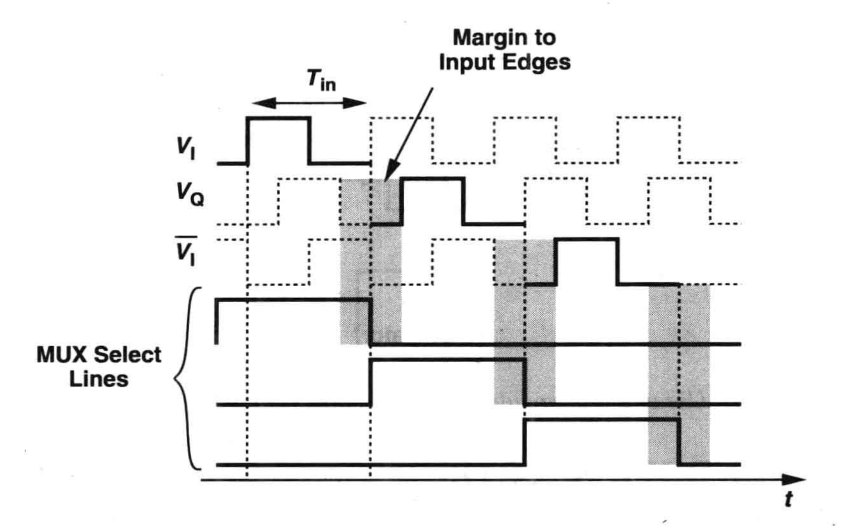


Figure 11.45 *Problem of phase selection timing margin.*

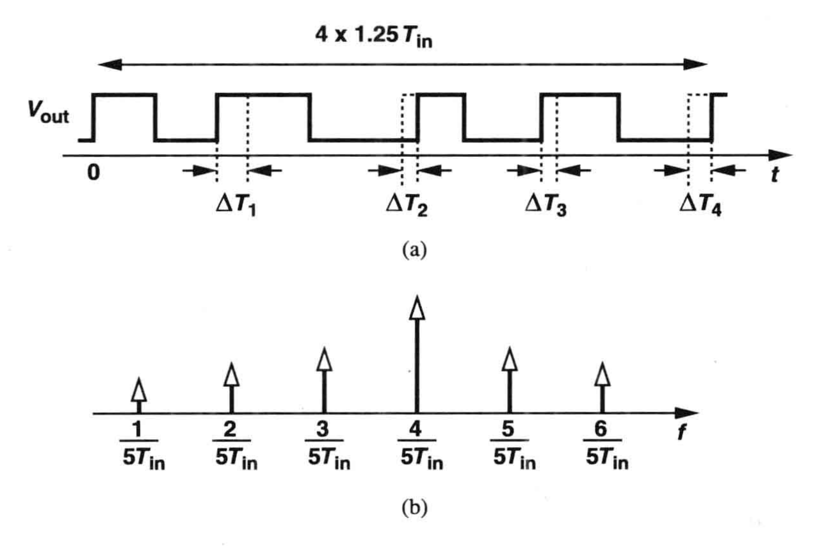


Figure 11.46 *(a) Effect of phase mismatches in VCO multiplexing, (b) resulting spectrum.*

points in time. As shown in Fig. 11.46(a), the consecutive periods are now unequal. Strictly speaking, we note that the waveform now repeats every $4 \times 1.25T_{in} = 5T_{in}$ seconds, exhibiting harmonics at $1/(5T_{in})$. That is, the spectrum contains a large component at $4/(5T_{in})$ and "sidebands" at other integer multiples of $1/(5T_{in})$ [Fig. 11.46(b)].

It is possible to randomize the selection of the phases in Fig. 11.44(a) so as to convert the sidebands to noise [10]. In fact, this randomization can incorporate noise shaping, leading to the architecture shown in Fig. 11.47. However, the first issue, namely, tight timing still remains. To relax this issue, the multiplexing of the VCO phases can be placed *after* the feedback divider [9, 10].

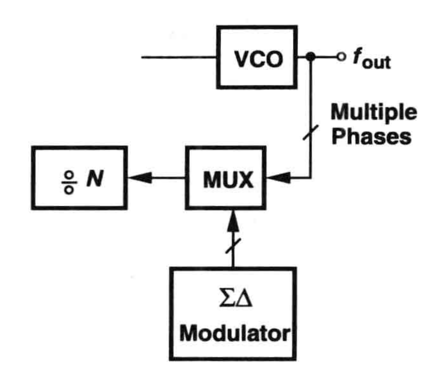


Figure 11.47 *Use of a ΣΔ modulator to randomize selection of the VCO phases.*

11.4 APPENDIX I: SPECTRUM OF QUANTIZATION NOISE

The random binary sequence, $b(t)$, in Fig. 11.6(a) consists of square pulses of width T_b that randomly repeat at a rate of $1/T_b$. In general, if a pulse $p(t)$ is randomly repeated every T_b seconds, the resulting spectrum is given by [11]:

$$S(f) = \frac{\sigma^2}{T_b}|P(f)|^2 + \frac{m^2}{T_b^2}\sum_{k=-\infty}^{+\infty}\left|P\left(\frac{k}{T_b}\right)\right|^2 \delta\left(f - \frac{k}{T_b}\right), \tag{11.57}$$

where σ^2 denotes the variance (power) of the data pulses, $P(f)$ the Fourier transform of $p(t)$, and m the mean amplitude of the data pulses. In our case, $p(t)$ is simply a square pulse toggling between 0 and 1 but with unequal probabilities: the probability that $p(t)$ occurs is the desired average, α (= m). The variance of a random variable x is obtained as

$$\sigma_x^2 = \int_{-\infty}^{+\infty} (x - m)^2 g(x) dx, \tag{11.58}$$

where $g(x)$ is the probability density function of x. For $b(t)$, $g(x)$ consists of an impulse of height $1 - \alpha$ at 0 and another height of α at 1 (why?) (Fig. 11.48). Thus,

$$\sigma^2 = \int_{-\infty}^{+\infty} \left[(0 - \alpha)^2(1 - \alpha)\delta(0) + (1 - \alpha)^2\alpha\delta(x - 1)\right] dx \tag{11.59}$$

$$= \alpha(1 - \alpha). \tag{11.60}$$

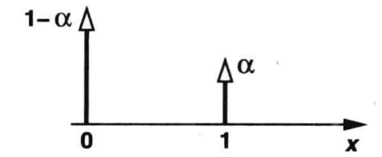


Figure 11.48 *Probability density function of binary data with an average value of α.*

Also, the Fourier transform of $p(t)$ is equal to

$$P(f) = \frac{\sin \pi f T_b}{\pi f}, \tag{11.61}$$

falling to *zero* at $f = k/T_b$ for $k \neq 0$. Thus, the second term in Eq. (11.57) reduces to $(\alpha^2/T_b)^2 |P(0)|^2 \delta(f) = \alpha^2 \delta(f)$. These derivations lead to Eq. (11.13).

REFERENCES

[1] T. A. D. Riley, M. A. Copeland, and T. A. Kwasniewski, "Delta-Sigma Modulation in Fractional-N Frequency Synthesis," *IEEE J. Solid-State Circuits,* vol. 28, pp. 553–559, May 1993.

[2] R. Schreier and G. C. Temes, *Understanding Delta-Sigma Data Converters,* New York: Wiley, 2004.

[3] P. Larsson, "High-Speed Architecture for a Programmable Frequency Divider and a Dual-Modulus Prescaler," *IEEE J. Solid-State Circuits,* vol. 31, pp. 744–748, May 1996.

[4] C. S. Vaucher et al., "A Family of Low-Power Truly Modular Programmable Dividers in Standard 0.35-μm CMOS Technology," *IEEE J. Solid-State Circuits,* vol. 35, pp. 1039–1045, July 2000.

[5] S. Pamarti, L. Jansson, and I. Galton, "A Wideband 2.4 GHz Delta-Sigma Fractional-N PLL with 1 Mb/s In-Loop Modulation," *IEEE J. of Solid State Circuits,* vol. 39, pp. 49–62, January 2004.

[6] S. E. Meninger and M. H. Perrott, "A 1-MHz Bandwidth 3.6-GHz 0.18-μm CMOS Fractional-N Synthesizer Utilizing a Hybrid PFD/DAC Structure for Reduced Broadband Phase Noise," *IEEE J. Solid-State Circuits,* vol. 41, pp. 966–981, April 2006.

[7] M. Gupta and B.-S. Song, "A 1.8-GHz Spur-Cancelled Fractional-N Frequency Synthesizer with LMS-Based DAC Gain Calibration," *IEEE J. Solid-State Circuits, ISSCC Dig. Tech. Papers,* pp. 323–324, Feb. 2006.

[8] H. Huh et al., "A CMOS Dual-Band Fractional-N Synthesizer with Reference Doubler and Compensated Charge Pump," *ISSCC Dig. Tech. Papers,* pp. 186–187, Feb. 2004.

[9] C.-H. Park, O. Kim, and B. Kim, "A 1.8-GHz Self-Calibrated Phase-Locked Loop with Precise I/Q Matching," *IEEE J. Solid-State Circuits,* vol. 36, pp. 777–783, May 2001.

[10] C.-H. Heng and B.-S. Song, "A 1.8-GHz CMOS Fractional-N Frequency Synthesizer with Randomized Multiphase VCO," *IEEE J. Solid-State Circuits,* vol. 38, pp. 848–854, June 2003.

[11] L. W. Couch, *Digital and Analog Communication Systems,* Fourth Edition, New York: Macmillan Co., 1993.

PROBLEMS

11.1. In the circuit of Fig. 11.1, $N = 10$, and $b(t)$ is a periodic waveform with $\alpha = 0.1$. Determine the spectrum of $f_{FB}(t) \approx (f_{out}/N)[1 - b(t)/N]$. Also, plot $q(t)$.

11.2. Based on the results from the previous problem, express the output phase of the divider as a function of time.

11.3. Suppose in Eq. (11.14), T_b is equal to the PLL input reference period. Recall that the loop bandwidth is about one-tenth of the reference frequency. What does this imply about the critical part of $S_q(f)$?

11.4. Extend the analysis leading to Eq. (11.42) for a third-order $\Sigma\Delta$ modulator and study the problem of out-of-band noise in this case.

11.5. Determine the noise-shaping function for a fourth-order $\Sigma\Delta$ modulator and compare its peak with that of a second-order modulator. For a given PLL bandwidth, how many more decibels of phase noise peaking does a fourth-order modulator create than a second-order counterpart?

11.6. Extend the approach illustrated in Fig. 11.22(b) to a cascade of a second-order system and a first-order system. Determine the logical operation of the output combiner.

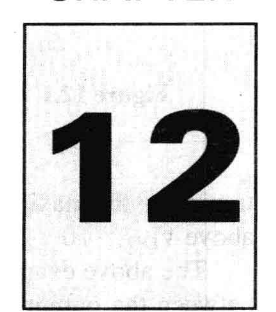
11.7. Suppose the Up and Down currents incur a 5% mismatch. Estimate the value of a in Eq. (11.45).

11.8. Determine whether two other effects in the PFD/CP combination result in noise folding: (a) unequal Up and Down pulsewidths, and (b) charge injection mismatch between the Up and Down switches in the charge pump.

11.9. Analyze the circuit of Fig. 11.38(a) if the MUX is driven by CK rather than $\overline{CK}$.

11.10. Show that the sideband at $1/(2T_1)$ in Fig. 11.42(b) is attenuated by approximately 16 dB in a PLL having $f_{out} = f_{in}$. What happens to the sideband magnitude if $f_{out} = Nf_{in}$? (Assume ζ and ω_n remain unchanged.)

CHAPTER



POWER AMPLIFIERS

Power amplifiers are the most power-hungry building block of RF transceivers and pose difficult design challenges. In the past ten years, the design of PAs has evolved considerably, drawing upon relatively complex transmitter architectures to improve the trade-off between linearity and efficiency. This chapter describes the analysis and design of PAs with particular attention to the limitations that they impose on the transmitter chain. A thorough treatment of PAs would require a book of its own, but our objective here is to lay the foundation. The reader is referred to [1, 2] for further details. The chapter outline is shown below.

Basic PA Classes	High-Efficiency PAs	Linearization Techniques	PA Design Examples
Class A PAs	Class A PAs with Harmonic Enhancement	Feedforward	Cascode PAs
Class B PAs	Class E PAs	Cartesian Feedback	Positive-Feedback PAs
Class C PAs	Class F PAs	Predistortion	PAs with Power Combining
		Polar Modulation	Polar Modulation PAs
		Outphasing	Outphasing PAs
		Doherty PA	

12.1 GENERAL CONSIDERATIONS

As the first step in our study, we consider a transmitter delivering 1 W (+30 dBm) of power to a 50-Ω antenna. The peak-to-peak voltage swing, V_{pp}, at the antenna reaches 20 V and the peak current through the load, 200 mA. For a common-source (or common-emitter) stage to drive the load directly, the configurations shown in Figs. 12.1(a) and (b) require a supply voltage greater than V_{pp}. However, if the load is realized as an inductor [Fig. 12.1(c)], the drain ac voltage *exceeds* V_{DD}, even reaching $2V_{DD}$ (or higher). While allowing a lower supply voltage, the inductive load does not relax the "stress" on the

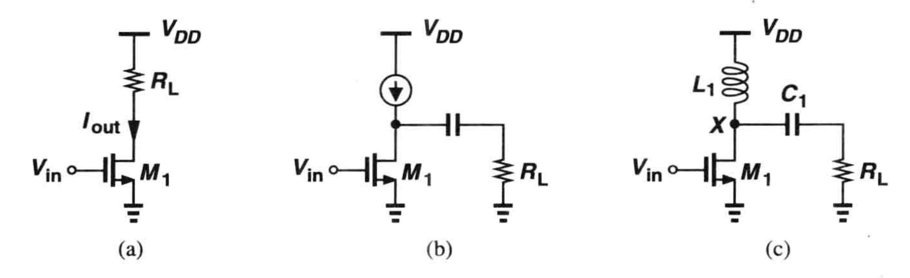


Figure 12.1 *CS stages with (a) resistive, (b) current source, and (c) inductive load.*

transistor; the maximum drain-source voltage experienced by M_1 is still at least 20 V (10 V above $V_{DD} = 10$ V) if the stage must deliver 1 W to a 50-Ω load.

The above example illustrates a fundamental issue in PA design, namely, the trade-off between the output power and the voltage swing experienced by the output transistor. It can be proven that the product of the breakdown voltage and f_T of silicon devices is around 200 GHz · V [3]. Thus, transistors with an f_T of 200 GHz dictate a voltage swing of less than 1 V.

Example 12.1

What is the peak current carried by M_1 in Fig. 12.1(c)? Assume L_1 is large enough to act as an ac open circuit at the frequency of interest, in which case it is called an "RF choke" (RFC).

Solution:

If L_1 is large, it carries a *constant* current, I_{L1} (why?). If M_1 begins to turn off, this current flows through R_L, creating a positive peak voltage of $I_{L1}R_L$ [Fig. 12.2(a)]. Conversely, if M_1 turns on completely, it must "sink" both the inductor current and a negative current of I_{L1} from R_L so as to create a peak voltage of $-I_{L1}R_L$ [Fig. 12.2(b)]. The peak current through the output transistor is therefore equal to 400 mA.

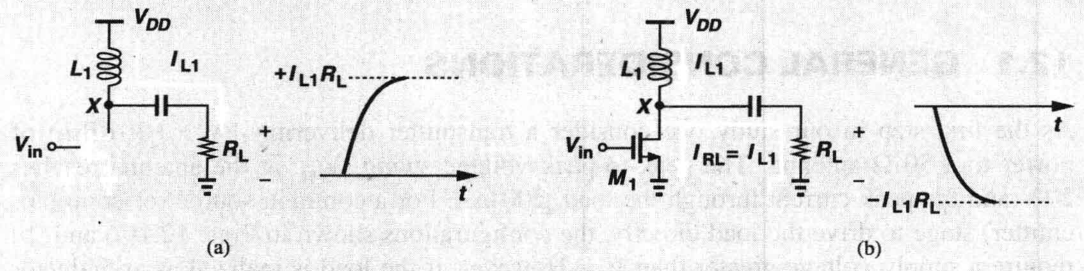


Figure 12.2 *Output voltage waveform in a CS stage (a) when current flows from inductor to R_L, (b) when current flows from R_L to transistor.*

In order to reduce the peak voltage experienced by the output transistor, a "matching network" is interposed between the PA and the load [Fig. 12.3(a)]. This network transforms

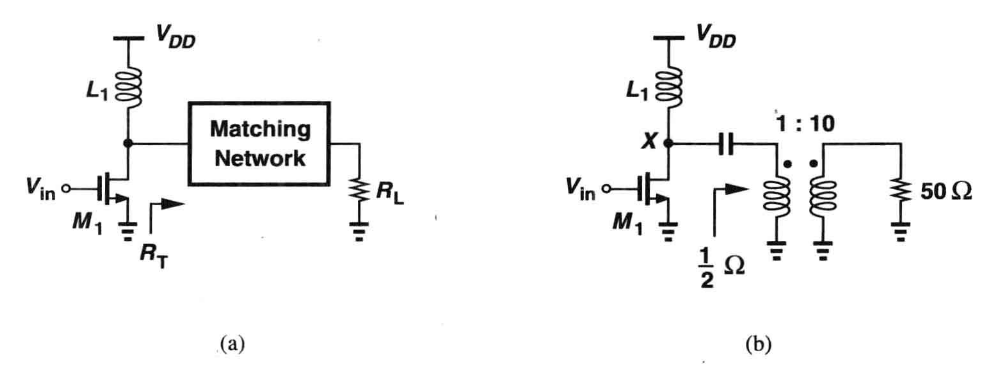


Figure 12.3 *(a) Impedance transformation by a matching network, (b) realization by a transformer.*

the load resistance to a *lower* value, R_T, so that smaller voltage swings still deliver the required power.

Example 12.2

The PA in Fig. 12.3(a) must deliver 1 W to $R_L = 50\ \Omega$ with a supply voltage of 1 V. Estimate the value of R_T.

Solution:

The peak-to-peak voltage swing, V_{pp}, at the drain of M_1 is approximately equal to 2 V. Since

$$P_{out} = \frac{1}{2}\left(\frac{V_{pp}}{2}\right)^2 \frac{1}{R_T} \tag{12.1}$$

$$= 1\ \text{W}, \tag{12.2}$$

we have

$$R_T = \frac{1}{2}\ \Omega. \tag{12.3}$$

The matching network must therefore transform R_L down by a factor of 100. Figure 12.3(b) shows an example, where a lossless transformer having a turns ratio of 1:10 converts a 2-V_{pp} swing at the drain of M_1 to a 20-V_{pp} swing across R_L.[1] From another perspective, the transformer amplifies the drain voltage swing by a factor of 10.

The need for transforming the voltage swings means that the current generated by the output transistor must be proportionally higher. In the above example, the peak current in the primary of the transformer reaches 10×200 mA $= 2$ A. Transistor M_1 must sink both the inductor current and the peak load current, i.e., 4 A!

1. A lossless transformer with a turns ratio of $1{:}n$ transforms the load resistance down by a factor of n^2 (why?).

Example 12.3

Plot V_X and V_{out} in Fig. 12.1(c) as a function of time if M_1 draws enough current to bring V_X near zero. Assume sinusoidal waveforms. Also, assume L_1 and C_1 are ideal and very large.

Solution:

In the absence of a signal, $V_X = V_{DD}$ and $V_{out} = 0$. Thus, the voltage across C_1 is equal to V_{DD}. We also observe that, in the steady state, the average value of V_X must be equal to V_{DD} because L_1 is ideal and therefore must sustain a zero average voltage. That is, if V_X goes from V_{DD} to near zero, it must also go from V_{DD} to about $2V_{DD}$ so that the average value of V_X is equal to V_{DD} (Fig. 12.4). The output voltage waveform is simply equal to V_X shifted down by V_{DD}.

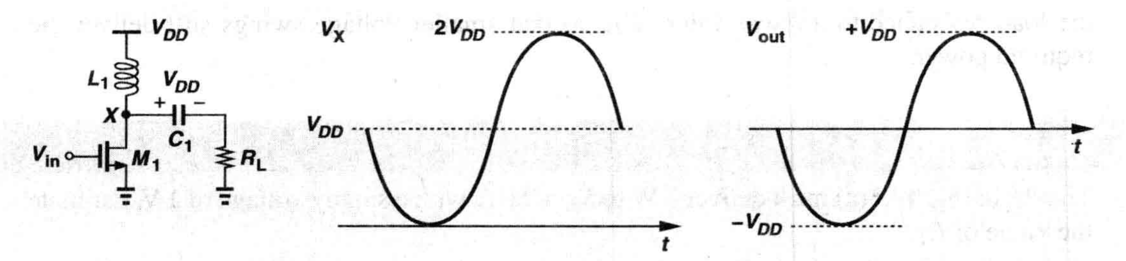


Figure 12.4 *Drain and output voltages in an inductively-loaded CS stage.*

12.1.1 Effect of High Currents

The enormous currents flowing through the output device and the matching network are one of the difficulties in the design of power amplifiers and the package. If the output transistor is chosen wide enough to carry a large current, then its *input* capacitance is very large, making the design of the *preceding* stage difficult. As depicted in Fig. 12.5, we may deal with this issue by interposing a number of tapered stages between the upconversion mixer(s) and the output stage. However, as explained in Chapter 4, the multiple stages tend to limit the TX output compression point. Moreover, the power consumed by the driver stages may not be negligible with respect to that of the output stage.

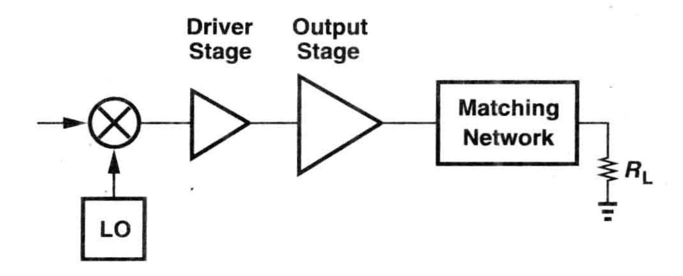


Figure 12.5 *Tapering in a TX chain.*

Another issue arising from the high ac currents in PAs relates to the package parasitics. The following example illustrates this point.

Example 12.4

The output transistor in Fig. 12.3(b) carries a current varying between 0 and 4 A at a frequency of 1 GHz. What is the maximum tolerable bond wire inductance in series with the source of the transistor if the voltage drop across this inductance must remain below 100 mV?

Solution:

The drain current of M_1 can be approximated as

$$I_D(t) = I_0 \cos \omega_0 t + I_0, \tag{12.4}$$

where $I_0 = 2$ A and $\omega_0 = 2\pi(1\text{ GHz})$. The voltage drop across the source inductance, L_S, is given by

$$V_{LS} = L_S \frac{dI_D}{dt}, \tag{12.5}$$

reaching a peak of $L_S \omega_0 I_0$. For this drop to remain below 100 mV, we have

$$L_S < 7.96\,\text{pH}. \tag{12.6}$$

This is an extremely small inductance. (A single bond wire's inductance typically exceeds 1 nH.)

What is the effect of package parasitics? The inductance in series with the source degenerates the transistor, thereby lowering the output power. Moreover, ground and supply inductances may create feedback from the output to the input of the PA chain, causing ripple in the frequency response and even instability.

The large currents can also lead to a high loss in the *matching network*. The devices comprising this network—especially the inductors—suffer from parasitic resistances, thus converting the signal energy to heat. For this reason, the matching network for high-power applications is typically realized with off-chip low-loss components.

12.1.2 Efficiency

Since PAs are the most power-hungry block in RF transceivers, their efficiency is critical. A 1-W PA with 50% efficiency draws 2 W from the battery—much more than the rest of the transceiver does.

The efficiency of the PAs is defined by two metrics. The "drain efficiency" (for FET implementations) or "collector efficiency" (for bipolar implementations) is defined as

$$\eta = \frac{P_L}{P_{supp}}, \tag{12.7}$$

where P_L denotes the average power delivered to the load and P_{supp} the average power drawn from the supply voltage. In some cases, the output stage may have a relatively *low* power gain, e.g., 3 dB, requiring a high *input* power. A quantity embodying this effect is the "power-added efficiency" (PAE), defined as

$$\text{PAE} = \frac{P_L - P_{in}}{P_{supp}}, \tag{12.8}$$

where P_{in} is the average input power.

Example 12.5

Discuss the PAE of the CS stage shown in Fig. 12.3.

Solution:

At low to moderate frequencies, the input impedance is capacitive and hence the average input power is zero. (Of course, driving a large capacitance is still difficult.) Thus, $\text{PAE} = \eta$. At high frequencies, the feedback due to the gate-drain capacitance introduces a real part in Z_{in}, causing the input port to draw some power.[2] Consequently, $\text{PAE} < \eta$. In stand-alone PAs, we may deliberately introduce a 50-Ω input resistance, in which case $\text{PAE} < \eta$.

12.1.3 Linearity

As explained in Chapter 3, the linearity of PAs becomes critical for some modulation schemes. In particular, PA nonlinearity leads to two effects: (1) high adjacent channel power as a result of spectral regrowth, and (2) amplitude compression. For example, QPSK modulation with baseband pulse shaping may suffer from the former and 16QAM from the latter. In some cases, AM/PM conversion may also be problematic.

The PA nonlinearity must be characterized with respect to the modulation scheme of interest. However, circuit-level simulations with actual modulated inputs take a very long time if they must produce an output spectrum that accurately reveals the ACPR (Chapter 3). Similarly, circuit-level simulations that quantify the effect of amplitude compression (i.e., the bit error rate) prove very cumbersome. For this reason, the PA characterization begins with two generic tests of nonlinearity based on unmodulated tones: intermodulation and compression. If employing two sufficiently large tones, the former provides some indication of ACPR. The amplitude of the tones is chosen such that each main component at the output is 6 dB below the full power level, thus producing the maximum desired output voltage swing when the two tones add in-phase [Fig. 12.6(a)]. For compression, a single tone is applied and its amplitude gradually increases so as to determine the output 1-dB compression point [Fig. 12.6(b)].

The above tests yield a first-order estimate of the PA nonlinearity. However, a more rigorous characterization is eventually necessary. Since the PA contains many storage elements, its nonlinearity cannot be simply expressed as a polynomial. As explained in

2. At very high frequencies, the gate and channel resistances also contribute a real part.

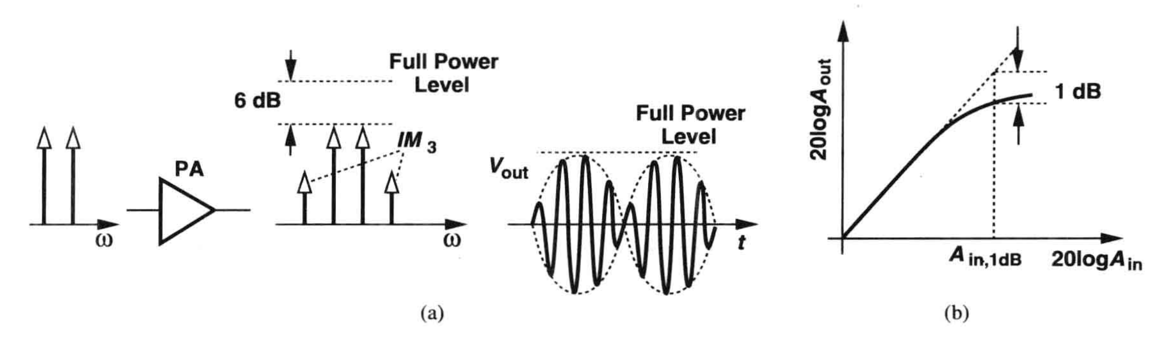


Figure 12.6 *PA characterization by (a) two-tone test, (b) compression.*

Chapter 2, a Volterra series can represent dynamic nonlinearities, but it tends to be rather complex. An alternative approach models the nonlinearity as follows [4]. Suppose the modulated input is of the form

$$x(t) = a(t)\cos[\omega_0 t + \phi(t)]. \tag{12.9}$$

Then, the output also contains amplitude and phase modulation and can be written as

$$y(t) = A(t)\cos[\omega_0 t + \phi(t) + \Theta(t)]. \tag{12.10}$$

We now make a "quasi-static" approximation. If the input signal bandwidth is much less than the PA bandwidth, i.e., if the PA can follow the signal dynamics closely, then we can assume that both $A(t)$ and $\Theta(t)$ are nonlinear static functions of only the input amplitude, $a(t)$. That is,

$$y(t) = A[a(t)]\cos\{\omega_0 t + \phi(t) + \Theta[a(t)]\}, \tag{12.11}$$

where $A[a(t)]$ and $\Theta[a(t)]$ represent "AM/AM conversion" and "AM/PM conversion," respectively [4]. For example, A and Θ are found to satisfy the following empirical equations:

$$A(a) = \frac{\alpha_1 a}{1 + \beta_1 a^2} \tag{12.12}$$

$$\Theta(a) = \frac{\alpha_2 a^2}{1 + \beta_2 a^2}, \tag{12.13}$$

where α_j and β_j are fitting parameters [4]. Illustrated in Fig. 12.7(a), $A(a)$ is similar to the characteristic shown in Fig. 12.6(b) (but declines for high input levels). The AM/PM conversion function can also be obtained relatively easily by applying a tone at the PA input and measuring the PA phase shift as a function of the input amplitude.

The reader may wonder why the foregoing model is valid. Indeed, no analytical proof appears to have been offered to justify this model. Nonetheless, it has been experimentally verified that the model provides reasonable accuracy if the input signal bandwidth remains much smaller than the PA bandwidth. Note that for a cascade of stages, the overall model may be quite complex and the behavior of A and Θ quite different.

With $A(a)$ and $\Theta(a)$ obtained from circuit simulations, the PA can be modeled by Eq. (12.11) and studied in a more efficient behavioral simulator, e.g., MATLAB. Thus, the

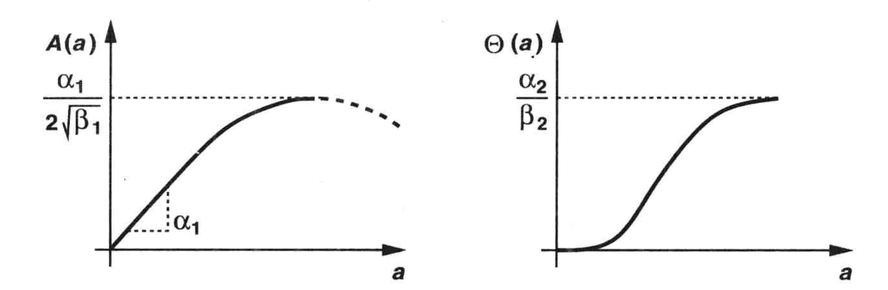


Figure 12.7 *Characteristics for AM/AM and AM/PM conversion.*

effect of the PA nonlinearity on ACPR or the quality of signals such as OFDM waveforms can be quantified.

Another PA nonlinearity representation, called the "Rapp model" [5], is expressed as follows:

$$g(V_{in}) = \frac{\alpha V_{in}}{[1 + (\frac{V_{in}}{V_0})^{2m}]^{\frac{1}{2m}}}, \tag{12.14}$$

where α denotes the small-signal gain around $V_{in} = 0$, and V_0 and m are fitting parameters. Dealing with only static nonlinearity, this model has become popular in integrated PA design. We return to this model in our back-off calculations in Chapter 13. Other PA modeling methods are described in [6].

12.1.4 Single-Ended and Differential PAs

Most stand-alone PAs have been designed as a cascade of single-ended stages. Two reasons account for this choice: the antenna is typically single-ended, and single-ended RF circuits are much simpler to *test* than their differential counterparts.

Single-ended PAs, however, suffer from two drawbacks. First, they "waste" half of the transmitter voltage gain because they sense only one output of the upconverter [Fig. 12.8(a)]. This issue can be alleviated by interposing a balun between the upconverter and the PA [Fig. 12.8(b)]. But the balun introduces its own loss, especially if it is integrated on the chip, limiting the voltage gain improvement to a few decibels (rather than 6 dB).

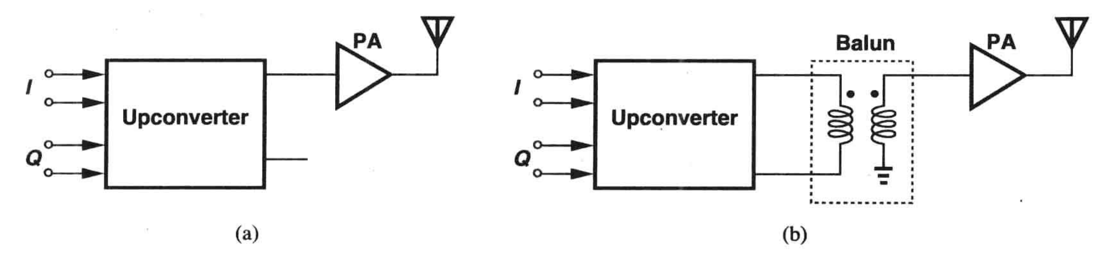


Figure 12.8 *Upconverter/PA interface with (a) single-ended or, (b) balun connection.*

The second drawback of single-ended PAs stems from the very large transient currents that they pull from the supply to the ground. As shown in Fig. 12.9(a), the supply bond wire inductance, L_{B1}, alters the resonance or impedance transformation properties of the output

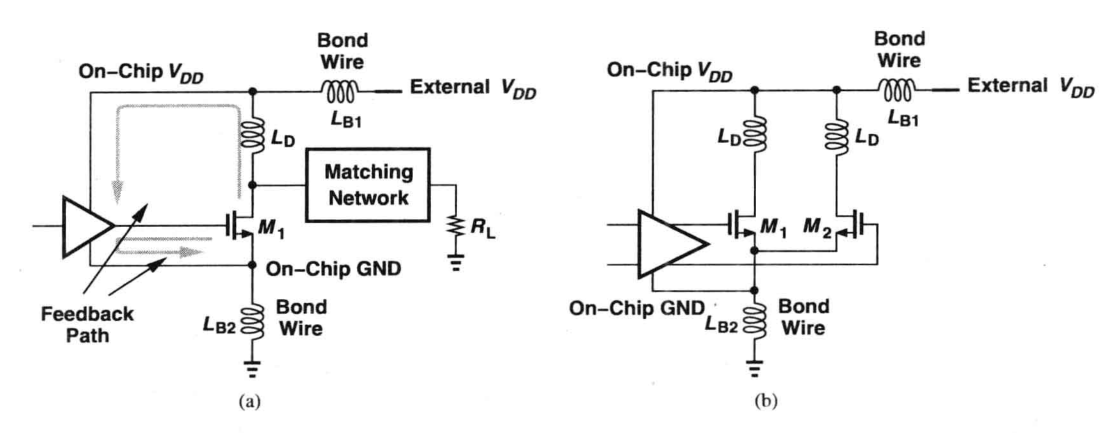


Figure 12.9 *(a) Feedback in a single-ended PA due to bond wires, (b) less problematic situation in a differential PA.*

network if it is comparable with L_D. Moreover, L_{B1} allows some of the output stage signal to travel back to the preceding stage(s) through the V_{DD} line, causing ripple in the frequency response or instability. Similarly, the ground bond wire inductance, L_{B2}, degenerates the output stage and introduces feedback.

By contrast, a differential realization greatly eases the above two issues. Illustrated in Fig. 12.9(b), such a topology draws much smaller transient currents from V_{DD} and ground lines, exhibiting less sensitivity to L_{B1} and L_{B2} and creating less feedback. The degeneration issue quantified in Example 12.4 is also relaxed considerably.

While the use of a differential PA ameliorates both the voltage gain and package parasitic issues, the PA must still drive a single-ended antenna in most cases. Thus, a balun must now be inserted between the PA and the antenna (Fig. 12.10).

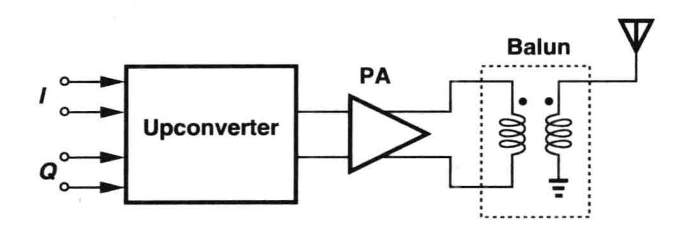


Figure 12.10 *Use of a balun between the PA and antenna.*

Example 12.6

Suppose a given balun design has a loss of 1.5 dB. In which one of the transmitters shown in Figs. 12.8(b) and 12.10 does this loss affect the efficiency more adversely?

Solution:

In Fig. 12.8(b), the balun lowers the voltage gain by 1.5 dB but does not consume much *power*. For example, if the power delivered by the upconverter to the PA is around 0 dBm,

(Continues)

Example 12.6 (*Continued*)

then a balun loss of 1.5 dB translates to a heat dissipation of 0.3 mW. In Fig. 12.10, on the other hand, the balun experiences the entire power delivered by the PA to the load, dissipating substantial power. For example, if the PA output reaches 1 W, then a balun loss of 1.5 dB corresponds to 300 mW. The TX efficiency therefore degrades more significantly in the latter case.

Another useful property of differential PAs is their lower coupling to the LO and hence reduced LO pulling (Chapter 4). If propagating symmetrically toward the LO, the differential waveforms generated by each stage of the PA tend to cancel. Of course, if the PA incorporates symmetric inductors, then the problem of coupling remains (Chapter 7).

The trade-offs governing the choice of single-ended and differential PAs has led to two schools of thought: some TX designs are based on fully-differential circuits with an on-chip or off-chip balun preceding the output matching network, while others opt for a single-ended PA—with or without a balun following the upconverter.

12.2 CLASSIFICATION OF POWER AMPLIFIERS

Power amplifiers have been traditionally categorized under many classes: A, B, C, D, E, F, etc. An attribute of classical PAs is that both the input and the output waveforms are considered sinusoidal. As we will see in Section 12.3, if this assumption is avoided, a higher performance can be achieved.

In this section, we describe classes A, B, and C, emphasizing their merits and drawbacks with respect to integrated implementation.

12.2.1 Class A Power Amplifiers

Class A amplifiers are defined as circuits in which the transistor(s) remain on and operate linearly across the full input and output range. Shown in Fig. 12.11 is an example. We note

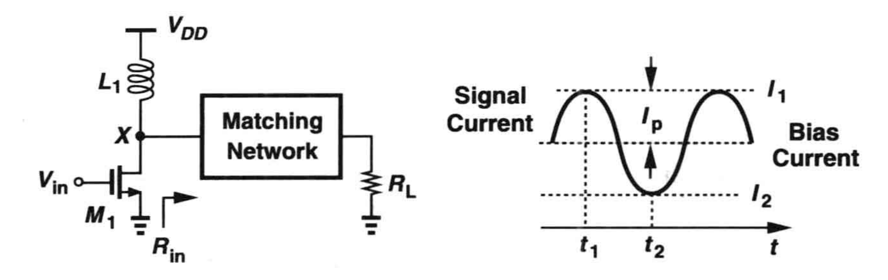


Figure 12.11 *Class A stage.*

that the transistor bias current is chosen higher than the peak signal current, I_p, to ensure that the device does not turn off at any point during the signal excursion.

The reader may wonder how we define "linear operation" here. After all, ensuring that the transistor is always on does not necessarily imply that the PA is sufficiently linear: if in Fig. 12.11, $I_1 = 5I_2$, the transistor transconductance varies considerably from t_1 to t_2 while the definition of class A seems to hold. This is where the definition of class A becomes vague. Nonetheless, we can still assert that *if* linearity is required, *then* class A operation is necessary.

Let us now compute the maximum drain (collector) efficiency of class A amplifiers. To reach maximum efficiency, we allow V_X in Fig. 12.11 to reach $2V_{DD}$ and nearly zero. Thus, the power delivered to the matching network is approximately equal to $(2V_{DD}/2)^2/(2R_{in}) = V_{DD}^2/(2R_{in})$, which is also delivered to R_L if the matching network is lossless. Also, recall from Example 12.1 that the inductive load carries a constant current of V_{DD}/R_{in} from the supply voltage. Thus,

$$\eta = \frac{V_{DD}^2/(2R_{in})}{V_{DD}^2/R_{in}} \tag{12.15}$$

$$= 50\%. \tag{12.16}$$

The other 50% of the supply power is dissipated by M_1 itself.

Example 12.7

Is the foregoing calculation of efficiency consistent with the assumption of linearity in class A stages?

Solution:

No, it is not. With a sinusoidal input, V_X in Fig. 12.11 reaches $2V_{DD}$ only if the transistor turns *off*. This ensures that the current swing delivered to the load goes from zero to twice the bias value.

It is important to recognize the assumptions leading to an efficiency of 50% in class A stages: (1) the drain (collector) peak-to-peak voltage swing is equal to *twice* the supply voltage, i.e., the transistor can withstand a drain-source (or collector-emitter) voltage of $2V_{DD}$ with no reliability or breakdown issues;[3] (2) the transistor barely turns off, i.e., the nonlinearity resulting from the very large change in the transconductance of the device is tolerable; (3) the matching network interposed between the output transistor and the antenna is lossless.

3. With a large voltage swing, the transistor may also introduce significant nonlinearity.

Example 12.8

Explain why low-gain output stages suffer from a more severe efficiency-linearity trade-off.

Solution:

Consider the two scenarios depicted in Fig. 12.12. In both cases, for M_1 to remain in saturation at $t = t_1$, the drain voltage must exceed $V_0 + V_{p,in} - V_{TH}$. In the high-gain stage of Fig. 12.12(a), $V_{p,in}$ is small, allowing V_X to come closer to zero than in the low-gain stage of Fig. 12.12(b).

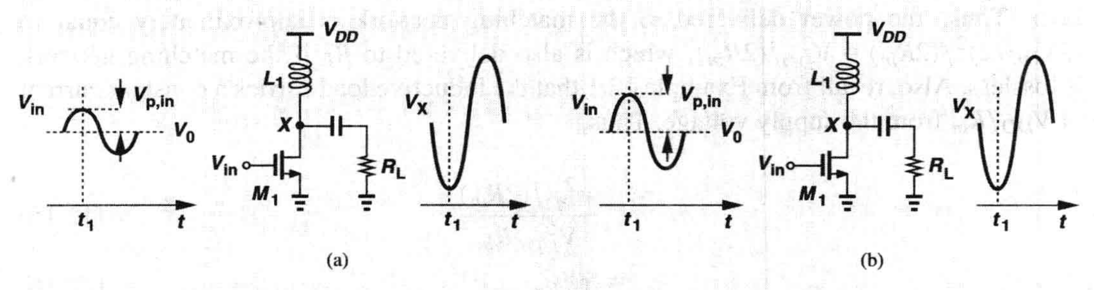


Figure 12.12 *Nonlinearity in a (a) high-gain and (b) low-gain stage.*

The above example indicates that the minimum drain voltage may *not* be negligible with respect to V_{DD}, yielding an output swing less than $2V_{DD}$. We must therefore compute the efficiency for lower output signal levels. The result also proves useful in transmitters with a *variable* output power. For example, we note from Chapter 4 that CDMA networks require that the mobile continually adjust its transmitted power so that the base station receives an approximately constant level.

Suppose the PA in Fig. 12.11 must deliver a peak voltage swing of V_p to R_{in}, i.e., a power of $V_p^2/(2R_{in})$ to the antenna if the matching network is lossless. We consider three cases: (1) the supply voltage and bias current remain at the levels necessary for full output power $[V_{DD}^2/(2R_{in})]$ and only the input signal swing is reduced; (2) the supply voltage remains unchanged but the bias current is reduced in proportion to the output voltage swing; (3) both the supply voltage and the bias current are reduced in proportion to the output voltage swing.

In the first case, the bias current is equal to V_{DD}/R_{in} hence and a power of V_{DD}^2/R_{in} is drawn from the battery. Consequently,

$$\eta_1 = \frac{V_p^2/(2R_{in})}{V_{DD}^2/R_{in}} \tag{12.17}$$

$$= \frac{V_p^2}{2V_{DD}^2}. \tag{12.18}$$

The efficiency thus falls sharply as the input and output voltage swings decrease.

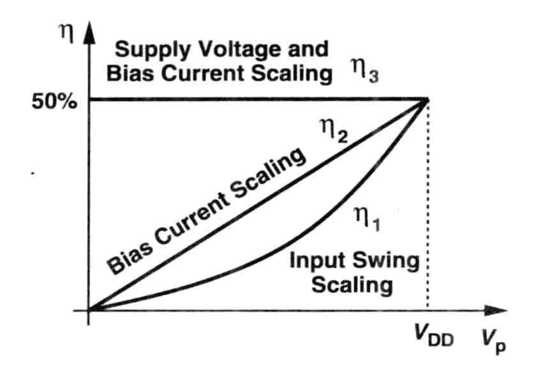


Figure 12.13 *Efficiency as a function of peak output voltage for different scaling scenarios.*

In the second case, the bias current is reduced to that necessary for a peak swing of V_p, i.e., V_p/R_{in}. It follows that

$$\eta_2 = \frac{V_p^2/(2R_{in})}{(V_p/R_{in})V_{DD}} \tag{12.19}$$

$$= \frac{V_p}{2V_{DD}}. \tag{12.20}$$

Here, the efficiency falls linearly as V_p decreases and V_{DD} remains constant.

In the third case, the supply voltage is also scaled, ideally according to the relation $V_{DD} = V_p$. Thus,

$$\eta_3 = 50\%. \tag{12.21}$$

While this case is the most desirable, it is difficult to design PA stages with a variable supply voltage. Figure 12.13 summarizes the results.

Example 12.9

A student attempts to construct an output stage with a variable supply voltage as shown in Fig. 12.14. Here, M_2 operates in the triode region, acting as a voltage-controlled resistor, and C_2 establishes an ac ground at node Y. Can this circuit achieve an efficiency of 50%?

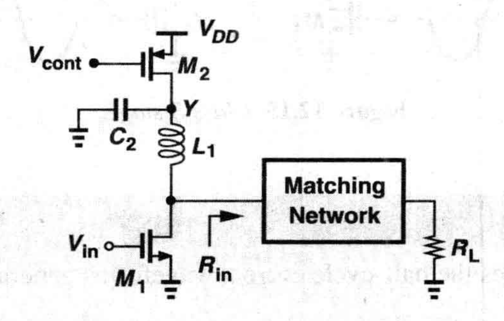


Figure 12.14 *Output stage with variable supply voltage.*

(Continues)

Example 12.9 (*Continued*)

Solution:

No, it cannot. Unfortunately, M_2 itself consumes power. If the bias current is chosen equal to V_p/R_{in}, then the total power drawn from V_{DD} is still given by $(V_p/R_{in})V_{DD}$ regardless of the on-resistance of M_2. Thus, M_2 consumes a power of $(V_p/R_{in})R_{on2}$, where R_{on2} denotes its on-resistance.

Conduction Angle It is sometimes helpful to distinguish PA classes by the "conduction angle" of their output transistor(s). The conduction angle is defined as the percentage of the signal period during which the transistor(s) remain on multiplied by 360°. In class A stages, the conduction angle is 360° because the output transistor is always on.

12.2.2 Class B Power Amplifiers

The definition of class B operation has changed over time! The traditional class B PA employs two *parallel* stages each of which conducts for only 180°, thereby achieving a higher efficiency than the class A counterpart. Figure 12.15 shows an example, where the drain currents of M_1 and M_2 are combined by transformer T_1. We may view the circuit as a quasi-differential stage and a balun driving the single-ended load. But class B operation requires that each transistor *turn off* for half of the period (i.e., the conduction angle is 180°). The gate bias voltage of the devices is therefore chosen approximately equal to their threshold voltage.

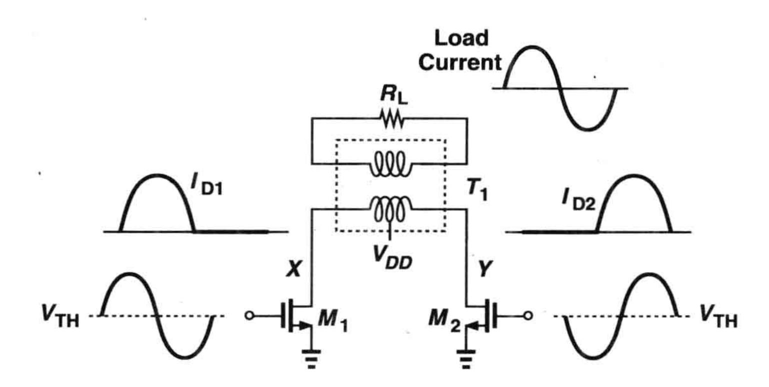


Figure 12.15 *Class B stage.*

Example 12.10

Explain how T_1 combines the half-cycle current waveforms generated by M_1 and M_2.

Solution:

Using superposition, we draw the output network in the two half cycles as shown in Fig. 12.16. When M_1 is on, I_{D1} flows from node X, producing a current in the secondary

Example 12.10 *(Continued)*

that flows *into* R_L and generates a positive V_{out} [Fig. 12.16(a)]. Conversely, when M_2 is on and draws current from node Y, the secondary current flows *out of* R_L and generates a negative V_{out} [Fig. 12.16(b)].

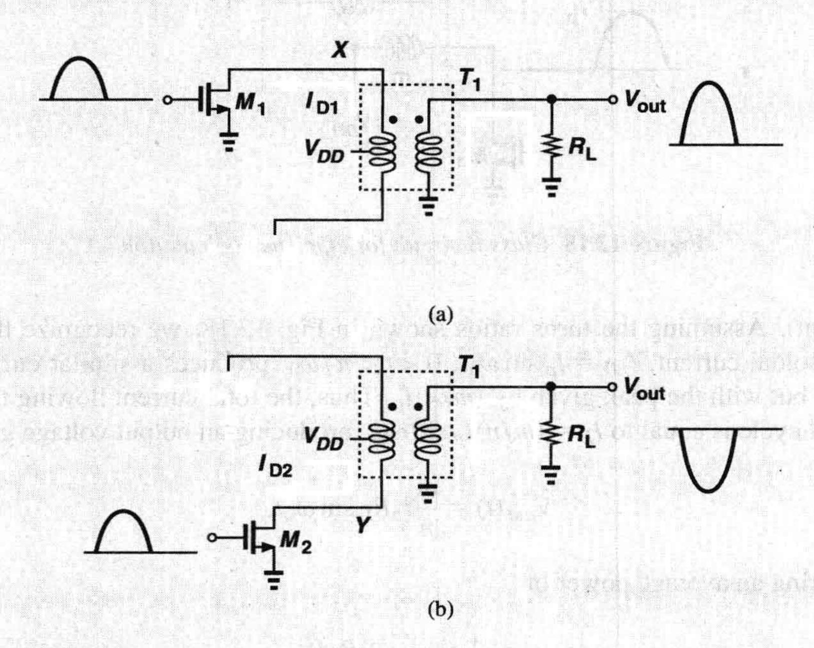


Figure 12.16 *Output network currents during (a) positive and (b) negative output half cycles.*

If the parasitic capacitances are small and the primary and secondary inductances are large, then V_X and V_Y in Fig. 12.15 are also half-wave rectified sinusoids that swing around V_{DD} (Fig. 12.17). In Problem 12.3, we show that the swing above V_{DD} is approximately half that below V_{DD}, an undesirable situation because it results in a low efficiency. For this reason, the secondary (or primary) of the transformer is tuned by a parallel capacitance so as to suppress the harmonics of the half-wave rectified sinusoids at X and Y, allowing equal swings above and below V_{DD}.

Let us compute the efficiency of the class B stage shown in Fig. 12.15. Suppose each transistor draws a peak current of I_p from the primary. As explained in Example 12.10, this current flows through *half* of the primary winding (because the other half carries a

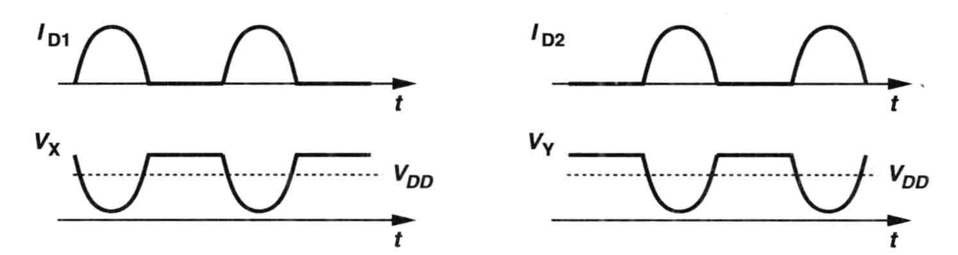


Figure 12.17 *Current and voltage waveforms in a class B stage.*

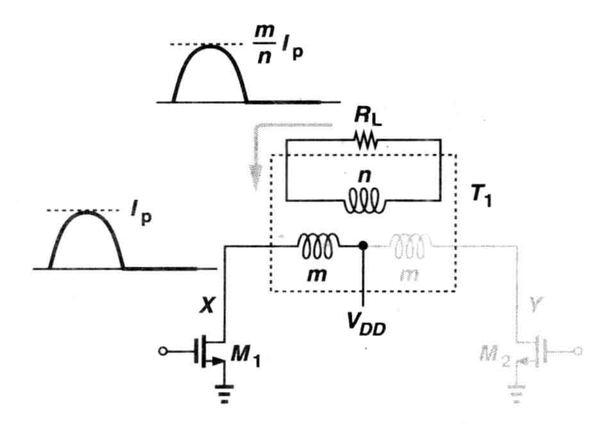


Figure 12.18 *Class B circuit for efficiency calculation.*

zero current). Assuming the turns ratios shown in Fig. 12.18, we recognize that a half-cycle sinusoidal current, $I_{D1} = I_p \sin \omega_0 t$, $0 < t < \pi/\omega_0$, produces a similar current in the secondary, but with the peak given by $(m/n)I_p$. Thus, the total current flowing through R_L in each full cycle is equal to $I_L = (m/n)I_p \sin \omega_0 t$, producing an output voltage given by

$$V_{out}(t) = \frac{m}{n} I_p R_L \sin \omega_0 t, \tag{12.22}$$

and delivering an average power of

$$P_{out} = \left(\frac{m}{n}\right)^2 \frac{R_L I_p^2}{2}. \tag{12.23}$$

We must now determine the average power drawn from V_{DD}. The half-wave rectified current drawn by each transistor has an average of I_p/π (why?). Since two of these current waveforms are drawn from V_{DD} in each period, the average power provided by V_{DD} is equal to

$$P_{supp} = 2\frac{I_p}{\pi} V_{DD}. \tag{12.24}$$

Dividing Eqs. (12.23) by (12.24) gives the drain (collector) efficiency of class B stages:

$$\eta = \frac{\pi}{4V_{DD}} \left(\frac{m}{n}\right)^2 I_p R_L. \tag{12.25}$$

As expected, η is a function of I_p.

In our last step, we calculate the voltage swings at X and Y in the presence of a resonant load in the secondary (or primary). Since the resonance suppresses the higher harmonics of the half-wave rectified cycles, V_X and V_Y resemble sinusoids that are 180° out of phase and have a dc level equal to V_{DD} (Fig. 12.19). That is,

$$V_X = V_p \sin \omega_0 t + V_{DD} \tag{12.26}$$

$$V_Y = -V_p \sin \omega_0 t + V_{DD}. \tag{12.27}$$

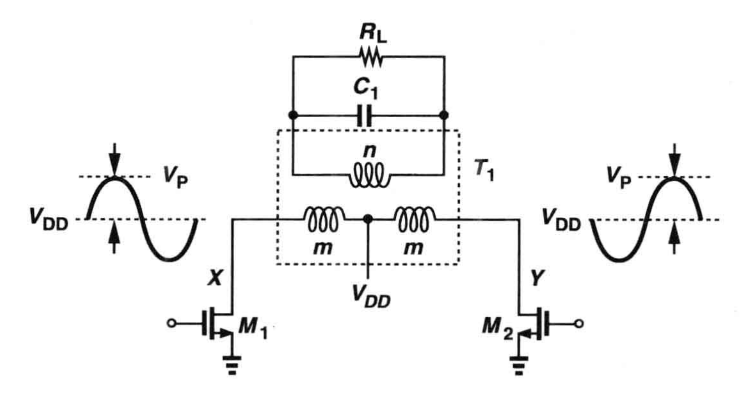


Figure 12.19 *Class B circuit with resonant secondary network.*

The primary of the transformer therefore senses a voltage waveform given by

$$V_{XY} = 2V_p \sin \omega_0 t, \tag{12.28}$$

which, upon experiencing a ratio of $n/(2m)$, yields the output voltage:

$$V_{out}(t) = \left(\frac{n}{2m}\right) 2V_p \sin \omega_0 t \tag{12.29}$$

$$= \frac{m}{n} I_p R_L \sin \omega_0 t. \tag{12.30}$$

It follows that

$$V_p = \frac{m^2}{n^2} I_p R_L. \tag{12.31}$$

We choose $V_p = V_{DD}$ to maximize the efficiency, obtaining from Eq. (12.25)

$$\eta = \frac{\pi}{4} \tag{12.32}$$

$$\approx 79\%. \tag{12.33}$$

In recent RF design literature, class B operation often refers to *half* of the circuits shown in Figs. 12.15 and 12.18, with the transistor still conducting for only half a cycle. Such a circuit, of course, is quite nonlinear but still has a maximum efficiency of $\pi/4$.

As mentioned in Section 12.1.4, the use of an on-chip balun at the PA output lowers the efficiency. For power levels above roughly 100 mW, an off-chip balun may be used if efficiency is critical.

Class AB Power Amplifiers The term "class AB" is sometimes used to refer to a single-ended PA (e.g., a CS stage) whose conduction angle falls between 180° and 360°, i.e., in which the output transistor turns off for less than half of a period. From another perspective, a class AB PA is less linear than a class A stage and more linear than a class B stage. This is usually accomplished by reducing the input voltage swing and hence backing off from the 1-dB compression point. Nonetheless, the term class AB remains vague.

12.2.3 Class C Power Amplifiers

Our study of class A and B stages indicates that a smaller conduction angle yields a higher efficiency. In class C stages, this angle is reduced further (and the circuit becomes more nonlinear).

The class A topology of Fig. 12.11 can be modified to operate in class C. Depicted in Fig. 12.20(a), the circuit is biased such that M_1 turns on if the peak value of V_{in} raises V_X above V_{TH}. As illustrated in Fig. 12.20(b), V_X exceeds V_{TH} for only a fraction of the period, as if M_1 were stimulated by a narrow pulse. As a result, the transistor delivers a narrow pulse of current to the output every cycle. In order to avoid large harmonic levels at the antenna, the matching network must provide some filtering. In fact, the input impedance of the matching network is also designed to resonate at the frequency of interest, thereby making the *drain voltage* a sinusoid.

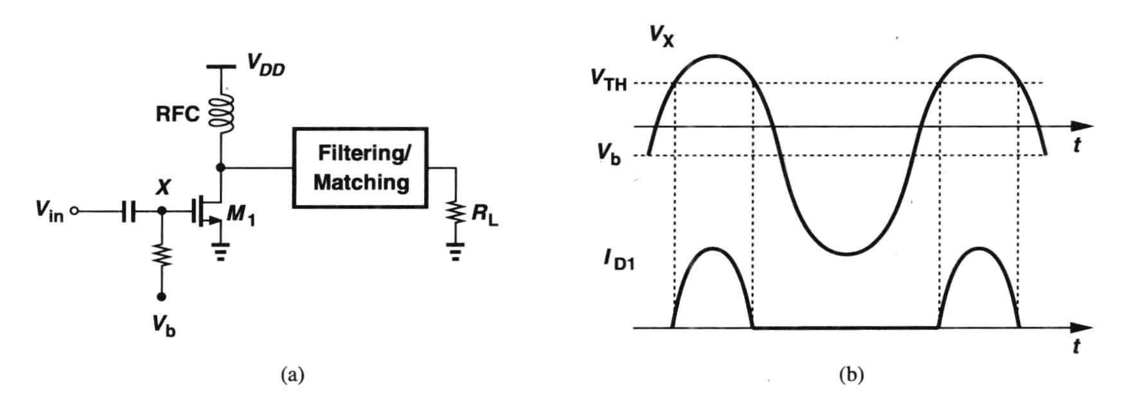


Figure 12.20 *(a) Class C stage and (b) its waveforms.*

The distinction between class C and one-transistor class B stages is in the conduction angle, θ. As θ decreases, the transistor is on for a smaller fraction of the period, thus dissipating less power. For the same reason, however, the transistor delivers less power to the load.

If the drain current of M_1 in Fig. 12.20(a) is assumed to be the peak section of a sinusoid and the drain voltage a sinusoid having a peak amplitude of V_{DD}, then the efficiency can be obtained as [7]

$$\eta = \frac{1}{4}\frac{\theta - \sin\theta}{\sin(\theta/2) - (\theta/2)\cos(\theta/2)}. \tag{12.34}$$

Sketched in Fig. 12.21(a), this relation suggests an efficiency of 100% as θ approaches zero.

The maximum efficiency of 100% is often considered a prominent feature of class C stages. However, another attribute that must also be taken into account is the actual power delivered to the load. It can be proved that [7]

$$P_{out} \propto \frac{\theta - \sin\theta}{1 - \cos(\theta/2)}. \tag{12.35}$$

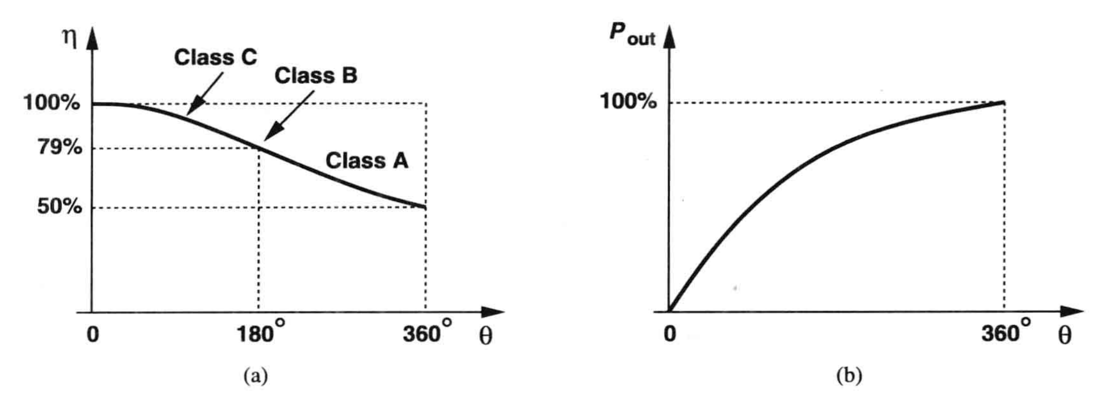


Figure 12.21 *(a) Efficiency and (b) output power as a function of conduction angle.*

Applying L'Hopital's rule, the reader can prove that P_{out} falls to zero as θ approaches zero. In other words, for a given design, a class C stage provides a high efficiency only if it delivers a *fraction* of the peak output power (the power corresponding to full class A operation).

How can a class C stage provide an output power comparable to that of a class A design? The small conduction angle dictates that the output transistor be *very wide* so as to deliver a high current for a short amount of time. In other words, the *first* harmonic of the drain current must be equal in the two cases.

Example 12.11

Determine the amplitude of the first harmonic of the transistor drain current in Fig. 12.20 for a conduction angle of θ.

Solution:

Consider the waveform shown in Fig. 12.22, where conduction begins at point A and ends at point B. The angle of the sinusoid reaches α at A and $\pi - \alpha$ at B such that $\pi - \alpha - \alpha = \theta$

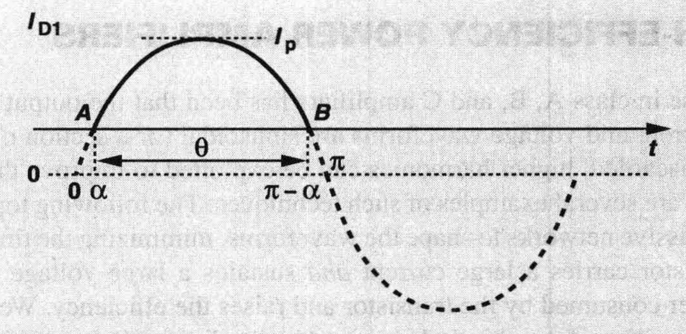


Figure 12.22 *Waveform in a class C stage for harmonic calculation.*

(Continues)

Example 12.11 *(Continued)*

and hence $\alpha = (\pi - \theta)/2$. The Fourier coefficients of the first harmonic are obtained as

$$a_1 = \frac{2}{T_0} \int_{\alpha/\omega 0}^{(\pi-\alpha)/\omega 0} I_p \sin \omega_0 t \sin \omega_0 t dt \tag{12.36}$$

$$b_1 = \frac{2}{T_0} \int_{\alpha/\omega 0}^{(\pi-\alpha)/\omega 0} I_p \sin \omega_0 t \cos \omega_0 t dt, \tag{12.37}$$

where $T_0 = 2\pi/\omega_0$ is the period. It follows that

$$a_1 = I_p \frac{\pi - 2\alpha}{2\pi} + \frac{I_p}{2\pi} \sin 2\alpha \tag{12.38}$$

$$b_1 = 0 \tag{12.39}$$

and hence the first harmonic is expressed as

$$I_{\omega 0}(t) = a_1 \sin \omega_0 t. \tag{12.40}$$

Note that $a_1 \rightarrow 0$ as $\alpha \rightarrow \pi/2$. For example, if $\alpha = \pi/4$, then $a_1 \approx 0.41 I_p$, the transistor must therefore be about 2.4 times as large as in a class-A stage for the same output power. Upon multiplication by R_{in}, this harmonic must yield a drain voltage swing of nearly $2V_{DD}$.

In modern RF design, class C operation has been replaced by other efficient amplification techniques that do not require such large transistors.

12.3 HIGH-EFFICIENCY POWER AMPLIFIERS

The main premise in class A, B, and C amplifiers has been that the output transistor drain (or collector) current and voltage waveforms are sinusoidal (or a section of a sinusoid). If this premise is discarded, higher harmonics can be exploited to improve the performance. Described below are several examples of such techniques. The following topologies rely on specific output passive networks to shape the waveforms, minimizing the time during which the output transistor carries a large current *and* sustains a large voltage. This approach reduces the power consumed by the transistor and raises the efficiency. We note, however, that the large parasitics of on-chip inductors typically dictate that matching networks be realized externally, making "fully-integrated PAs" a misnomer.

12.3.1 Class A Stage with Harmonic Enhancement

Recall from our study of the class A stage in Fig. 12.11 that, for maximum efficiency, the transistor current swings by a large amount, experiencing nonlinearity. Thus, the current contains a significant second and/or third harmonic. Now suppose the matching network is designed such that its input impedance is low at the fundamental and *high* at the second harmonic. As illustrated in Fig. 12.23, the sum of the resulting voltage waveforms exhibits narrower pulses than the fundamental, reducing the overlap time between the voltage across and the current flowing in the output transistor. Consequently, the average power consumed by the output transistor decreases and the efficiency increases.

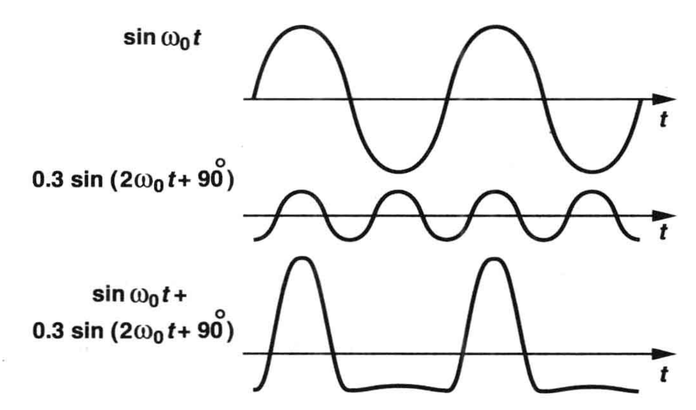


Figure 12.23 *Example of second harmonic enhancement.*

It is interesting that the above modification need not increase the harmonic content of the signal delivered to the load. The technique simply realizes different termination impedances for different harmonics to make the *drain* voltage approach a square wave.

As an example, consider the class A circuit shown in Fig. 12.24(a), where L_1, C_1 and C_2 form a matching network that transforms the 50-Ω load to $Z_1 = 9\ \Omega + j0$ at $f = 850$ MHz and $Z_2 = 330\ \Omega + j0$ at $2f = 1.7$ GHz [8]. In this case, the second harmonic is enhanced by a factor of 37. Figure 12.24(b) shows the drain voltage. The circuit delivers a power of 2.9 W to the load with 73% efficiency and a third-order harmonic of −25 dBc [8]. Other

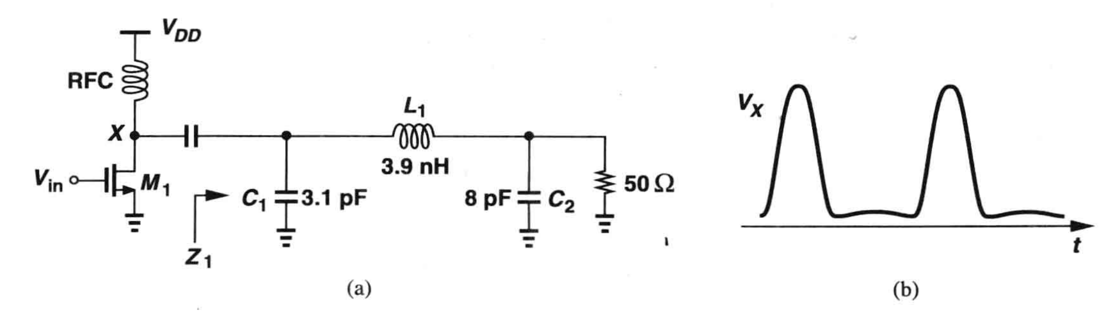


Figure 12.24 *(a) Class A stage with harmonic enhancement, (b) drain waveform.*

considerations for harmonic termination are described in [9]. This enhancement technique can be applied to other PA classes as well.

12.3.2 Class E Stage

Class E stages are nonlinear amplifiers that achieve efficiencies approaching 100% while delivering *full* power, a remarkable advantage over class C circuits. Before studying class E PAs in detail, we first revisit the simple circuit of Fig. 12.3(a), shown in Fig. 12.25.

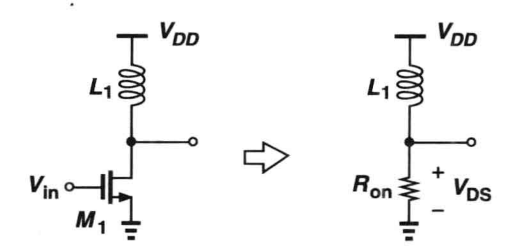


Figure 12.25 *Output stage with switching transistor.*

Suppose the output transistor in this circuit operates as a switch, rather than a voltage-dependent current source, ideally turning on and off abruptly. Called a "switching power amplifier," such a topology achieves a high efficiency if (1) M_1 sustains a small voltage when it carries current, (2) M_1 carries a small current when it sustains a finite voltage, and (3) the transition times between the on and off states are minimized [10]. From (1) and (3), we conclude that the on-resistance of the switch must be very small and the voltage applied to the gate of M_1 must approximate a rectangular waveform. However, even with these two conditions, (2) may still be violated if M_1 turns on when V_X is high. Of course, in practice it is difficult to obtain sharp input transitions at high frequencies.

It is important to understand the fundamental difference between the PAs studied in previous sections and the switching stage of Fig. 12.25: in the former, the output matching network is designed with the assumption that the transistor operates as a current source, whereas in the latter, this assumption is not necessary. If the transistor is to remain a current source, then the minimum value of the drain voltage and the maximum value of the gate voltage must be precisely controlled such that the transistor does not enter the triode region. The minimum required drain-source voltage translates to a lower efficiency even if all of the devices and waveforms are ideal. By contrast, in switching amplifiers the drain voltage can approach zero (or even a somewhat negative value).

A serious dilemma in nonlinear PA design is that the gate of the output device must be switched as abruptly as possible so as to maximize the efficiency [Fig. 12.26(a)], but the large output transistor typically necessitates resonance at its gate, inevitably receiving a nearly sinusoidal waveform [Fig. 12.26(b)].

Class E amplifiers deal with the finite input and output transition times by proper *load* design. Shown in Fig. 12.27(a), a class E stage consists of an output transistor, M_1, a grounded capacitor, C_1, and a series network C_2 and L_1 [10]. Note that C_1 includes the junction capacitance of M_1 and the parasitic capacitance of the RFC. The values of C_1, C_2, L_1, and R_L are chosen such that V_X satisfies three conditions: (1) as the switch turns off V_X remains low long enough for the current to drop to zero, i.e., V_X and I_{D1} have nonoverlapping waveforms [Fig. 12.27(b)]; (2) V_X reaches zero just before the switch turns

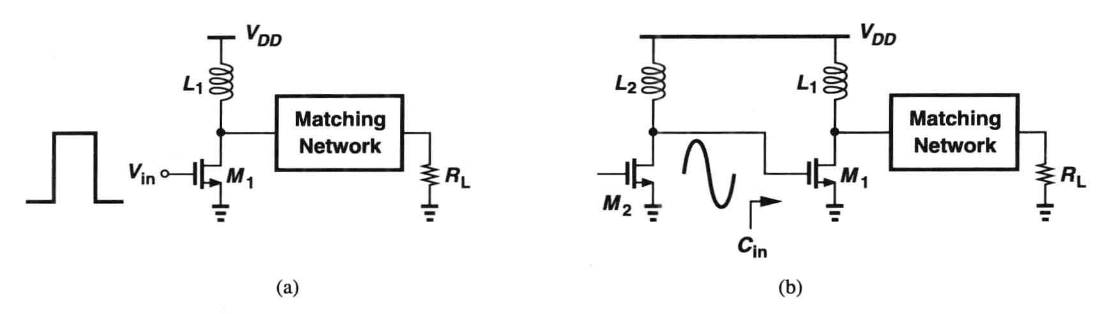


Figure 12.26 *(a) Switching stage with sharp input waveform, (b) gradual waveform due to resonance.*

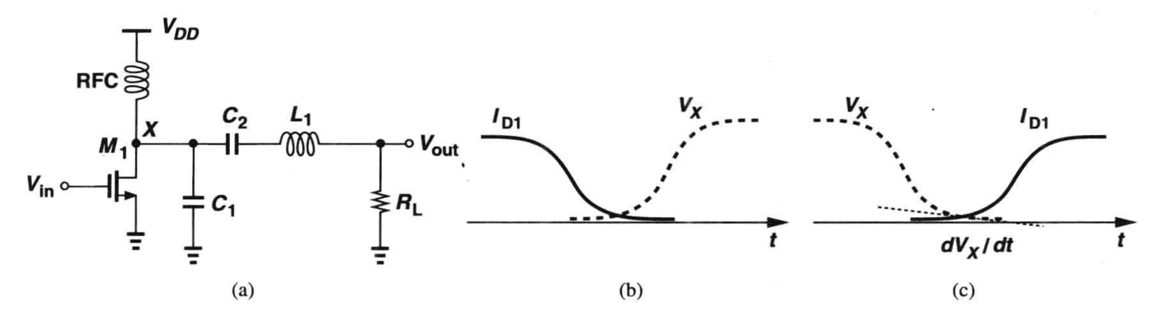


Figure 12.27 *(a) Class E stage, (b) condition to ensure minimal overlap between drain current and voltage, (c) condition to ensure low sensitivity to timing errors.*

on [Fig. 12.27(c)]; and (3) dV_X/dt is also near zero when the switch turns on. We examine these conditions to understand the circuit's properties.

The first condition, guaranteed by C_1, resolves the issue of finite fall time at the *gate* of M_1. Without C_1, V_X would rise as V_{in} dropped, allowing M_1 to dissipate substantial power.

The second condition ensures that the V_{DS} and I_D of the switching device do not overlap in the vicinity of the turn-on point, thus minimizing the power loss in the transistor even with finite input and output transition times.

The third condition lowers the sensitivity of the efficiency to violations of the second condition. That is, if device or supply variations introduce some overlap between the voltage and current waveforms, the efficiency degrades only slightly because $dV_X/dt = 0$ means V_X does not change significantly near the turn-off point.

The implementation of the second and third conditions is less straightforward. After the switch turns off, the load network operates as a damped second-order system (Fig. 12.28)

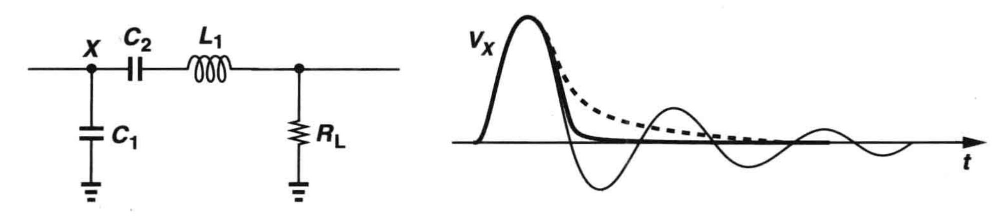


Figure 12.28 *Class E matching network viewed as a damped network.*

[10] with initial conditions across C_1 and C_2 and in L_1. The time response depends on the Q of the network and appears as shown in Fig. 12.28 for underdamped, overdamped, and critically-damped conditions. We note that in the last case, V_X approaches zero volt with zero slope. Thus, if the switch begins to turn on at this time, the second and third conditions are met.

Example 12.12

Modeling a class E stage as shown in Fig. 12.29(a), plot the circuit's voltages and currents.

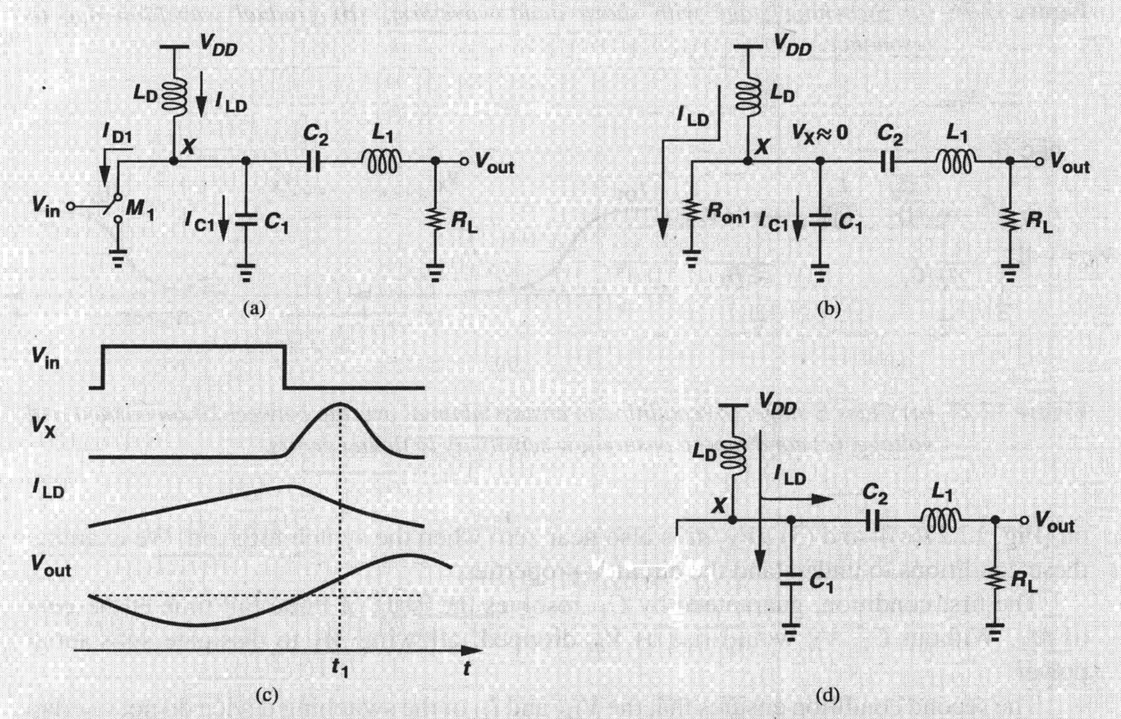


Figure 12.29 *(a) Model of class E stage, (b) simplified circuit when transistor is on, (c) voltage and current waveforms, (d) simplified circuit when transistor is off.*

Solution:

When M_1 turns on, it shorts node X to ground but carries little current because V_X is already near zero at this time (second condition described above) [Fig. 12.29(b)]. If R_{on1} is small, V_X remains near zero and L_D sustains a relatively constant voltage, thus carrying a current given by

$$I_{LD} = \frac{1}{L_D}\int (V_{DD} - V_X)dt \tag{12.41}$$

$$\approx \frac{V_{DD} - V_X}{L_D}t. \tag{12.42}$$

In other words, one half cycle is dedicated to charging L_D with minimal drop across M_1 [Fig. 12.29(c)]. When M_1 turns off, the inductor current begins to flow through C_1 and the

Example 12.12 *(Continued)*

load [Fig. 12.29(d)], raising V_X. This voltage reaches a peak at $t = t_1$ and begins to fall thereafter, approaching zero with a zero slope at the end of the second half cycle (second and third conditions described above). The matching network attenuates higher harmonics of V_X, yielding a nearly sinusoidal output.

Class E stages are quite nonlinear and exhibit a trade-off between efficiency and output harmonic content. For low harmonics, the Q of the output network must be higher than that typically required by the second and third conditions. In most standards, the harmonics of the carrier must be sufficiently small because they fall into other communication bands. (Note that a low harmonic content does not necessarily mean that the PA itself is linear; the output transistor may still create spectral regrowth or amplitude compression.)

Another property of class E amplifiers is the large peak voltage that the switch sustains in the off state, approximately $3.56V_{DD} - 2.56V_S$, where V_S is the minimum voltage across the transistor [10]. With $V_{DD} = 1$ V and $V_S = 50$ mV, the peak exceeds 3 V, raising serious device reliability or breakdown issues.

The design equations of class E stages are beyond the scope of this book. The reader is referred to [10] for details.

12.3.3 Class F Power Amplifiers

The idea of harmonic termination described in Section 12.3.1 can be extended to nonlinear amplifiers as well. If in the generic switching stage of Fig. 12.25 the load network provides a high termination impedance at the second or third harmonics, the voltage waveform across the switch exhibits sharper edges than a sinusoid, thereby reducing the power loss in the transistor. Such a circuit is called a class F stage [11].

Figure 12.30(a) shows an example of the class F topology. The tank consisting of L_1 and C_1 resonates at twice or three times the input frequency, approximating an open circuit. As depicted in Fig. 12.30(b), V_X approaches a rectangular waveform with the addition of the third harmonic.

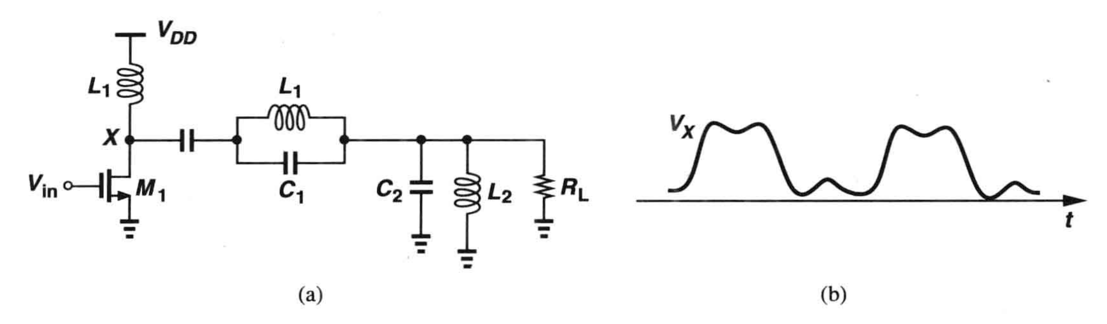


Figure 12.30 *Example of class F stage.*

Example 12.13

Explain why a class B stage does not lend itself to third-harmonic peaking.

Solution:

If the output transistor conducts for half of the cycle, the resulting half-wave rectified current contains *no* third harmonic. The Fourier coefficients of the third harmonic are given by

$$a_3 = \frac{1}{T_0}\int_0^{T_0/2} I_0 \sin\omega_0 t \sin 3\omega_0 t dt \tag{12.43}$$

$$= \frac{I_0}{2T_0}\int_0^{T_0/2} (\cos 2\omega_0 t - \cos 4\omega_0 t)dt \tag{12.44}$$

$$= 0 \tag{12.45}$$

and

$$b_3 = \frac{1}{T_0}\int_0^{T_0/2} I_0 \sin\omega_0 t \cos 3\omega_0 t dt \tag{12.46}$$

$$= \frac{I_0}{2T_0}\int_0^{T_0/2} (\sin 4\omega_0 t - \sin 2\omega_0 t)dt \tag{12.47}$$

$$= 0. \tag{12.48}$$

The above example suggests that third-harmonic peaking is viable only if the output transistor experiences "hard" switching, i.e., its output current resembles a rectangular wave. This in turn requires that the gate (or base) voltage be driven by relatively sharp edges.

If the drain current of the transistor is assumed to be a half-wave rectified sinusoid, it can be proved that the peak efficiency of class F amplifiers is equal to 88% for third-harmonic peaking [11].

12.4 CASCODE OUTPUT STAGES

Our study of PA stages in the previous sections reveals that to achieve a high efficiency, the output stage must produce a waveform that swings *above* V_{DD}. For example, in class

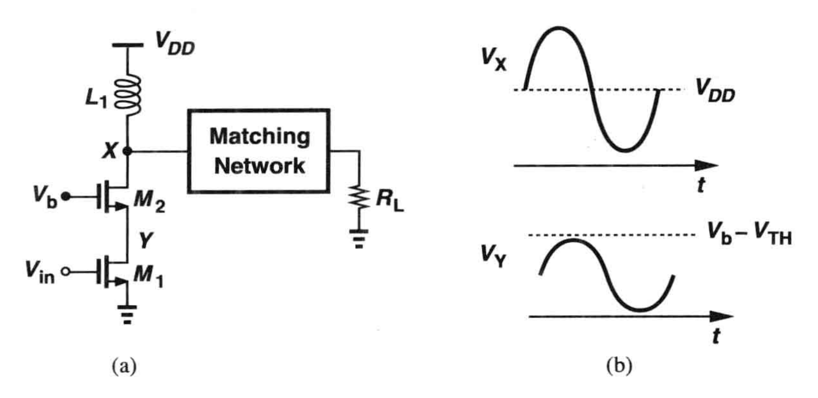


Figure 12.31 *(a) Cascode PA and (b) its waveforms.*

A and B efficiency calculations, the drain waveform is assumed to have a peak-to-peak swing of nearly $2V_{DD}$. However, if V_{DD} is chosen equal to the nominal supply voltage of the process, the output transistor experiences breakdown or substantial stress. One can choose V_{DD} equal to *half* of the maximum tolerable voltage of the transistor, but with two penalties: (a) the lower headroom limits the linear voltage range of the circuit, and (b) the proportionally higher output current (for a given output power) leads to a greater loss in the output matching network, reducing the efficiency.

A cascode output stage somewhat relaxes the above constraints. As shown in Fig. 12.31(a), the cascode device "shields" the input transistor as V_X rises, keeping the drain-source voltage of M_1 less than $V_b - V_{TH2}$ (why?). Depicted in Fig. 12.31(b) are the typical waveforms: V_X swings by about $2V_{DD}$ and V_Y by about $V_b - V_{TH}$ (if the minimum drain-source voltages are small).

Example 12.14

Determine the maximum terminal-to-terminal voltage differences of M_1 and M_2 in Fig. 12.31(a). Assume V_{in} has a peak amplitude of V_0 and a dc level of V_m, and V_X has a peak amplitude of V_p (and a dc level of V_{DD}).

Solution:

Transistor M_1 experiences maximum V_{DS} as V_{in} falls to $V_m - V_0$. If M_1 nearly turns off, then $V_{DS1} \approx V_b - V_{TH2}$, $V_{GS1} = V_m - V_0$, and $V_{DG1} = V_b - V_{TH2} - (V_m - V_0)$. For the same input level, the drain voltage of M_2 reaches its maximum of $V_{DD} + V_p$, creating

$$V_{DS2} = V_{DD} + V_p - (V_b - V_{TH2}), \tag{12.49}$$

and

$$V_{DG2} = V_{DD} + V_p - V_b. \tag{12.50}$$

Also, the drain-bulk voltage of M_2 reaches $V_{DD} + V_p$.

In the cascode topology of Fig. 12.31(a), the values of V_b and V_p must be chosen so as to guarantee V_{DS2} and V_{DG2} remain below V_{DD} at all times. (The drain-bulk voltage is typically allowed to reach $2V_{DD}$ or even higher with no reliability concerns.) From Eqs. (12.49) and (12.50), we can write respectively,

$$V_{DD} + V_p - V_b + V_{TH2} \le V_{DD} \tag{12.51}$$

$$V_{DD} + V_p - V_b \le V_{DD}. \tag{12.52}$$

The former is a stronger condition and reduces to

$$V_p \le V_b - V_{TH2}. \tag{12.53}$$

For example, if $V_b = V_{DD}$, then $V_p \le V_{DD} - V_{TH2}$; i.e., the peak-to-peak swing at X is limited to $2V_{DD} - 2V_{TH2}$. With body effect, V_{TH2} may reach 0.5 V in 90-nm and 65-nm technologies, yielding a total swing of only 1 V_{pp}, about the same as that of a non-cascoded common-source stage! We therefore observe that the cascode topology offers only a marginal increase in the maximum allowable output swing at low supply voltages.[4] Since a cascode topology with a supply voltage of V_{DD} provides an output swing approximately equal to that of a common-source stage with a supply voltage of $V_{DD}/2$, we expect the former to exhibit an efficiency about half that of the latter, i.e., about 25% in class A operation.

Let us now compare the cascode and CS stages in terms of their linearity. For the stages shown in Fig. 12.32, we seek the maximum output voltage swing that places M_1 at the edge of saturation. From Fig. 12.32(a),

$$V_{DD} - V_{p,cas} - V_{DS2} + V_{TH1} = V_0 + V_m, \tag{12.54}$$

and from Fig. 12.32(b),

$$V_{DD} - V_{p,CS} + V_{TH1} = V_0 + V_m. \tag{12.55}$$

It follows that

$$V_{p,CS} = V_{p,cas} + V_{DS2}. \tag{12.56}$$

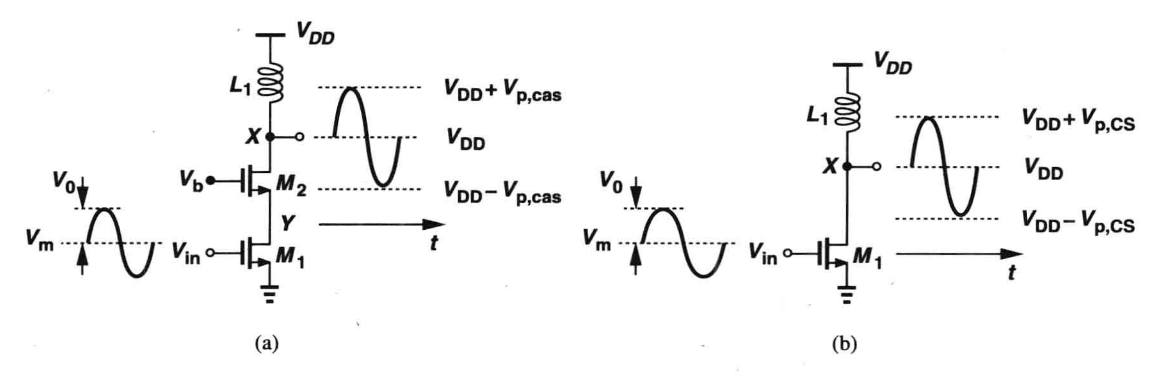


Figure 12.32 *(a) Cascode and (b) CS stages for linearity analysis.*

4. This issue can be alleviated through the use of a low-threshold transistor for M_2.

Thus, the CS stage remains linear across a wider output voltage range than the cascode circuit does.

The foregoing study suggests that, at low supply voltages, cascode output stages offer only a slight voltage swing advantage over their CS counterparts, but at the cost of efficiency and linearity. Nonetheless, by virtue of their high reverse isolation (a small $|S_{12}|$), cascode stages experience less feedback, thus proving more stable. As studied in Chapter 5 for low-noise amplifiers, a simple CS stage may suffer from a negative input resistance.

Example 12.15

Consider the two-stage PA shown in Fig. 12.33(a). If the output stage exhibits a negative input resistance, how can the cascade be designed to remain stable?

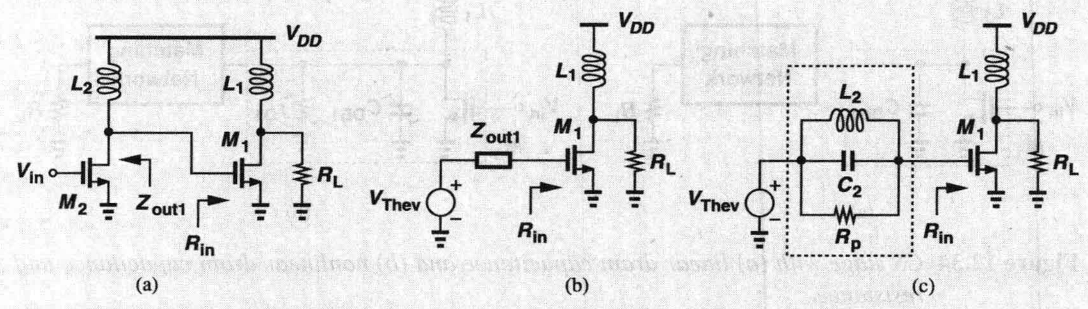


Figure 12.33 *(a) Cascade of two CS stages, (b) simplified model of (a), (c) representation of first stage by a resonant impedance.*

Solution:

Drawing the Thevenin equivalent of the first stage as shown in Fig. 12.33(b), we observe that instability can be avoided if

$$Re\{Z_{out1}\} + R_{in} > 0 \tag{12.57}$$

so that V_{Thev} does not *absorb* energy from the circuit. If Z_{out} is modeled by a parallel tank [Fig. 12.33(c)], then

$$Re\{Z_{out1}\} = R_p. \tag{12.58}$$

Thus, we require that

$$R_p + R_{in} > 0. \tag{12.59}$$

Of course, this condition must hold at all frequencies and for a certain range of R_{in}. For example, if the user of a cell phone wraps his/her hand around the antenna, R_L and hence R_{in} change.

We deal with the transistor-level design of a 6-GHz cascode PA in Chapter 13. The efficiency of the circuit reaches 30% around compression but falls to 5% with enough back-off to satisfy 11a requirements.

12.5 LARGE-SIGNAL IMPEDANCE MATCHING

In the development of PAs thus far, we have assumed that the output matching network simply transforms R_L to a lower value. This simplistic model of the output network is shown in Fig. 12.34(a), where M_1 operates as an ideal current source and L_1 resonates with C_{DB1}, allowing the transistor's RF current to flow into R_L. In practice, however, the situation is more complex: the transistor exhibits an output resistance, r_{O1}, and both r_{O1} and C_{DB1} *vary* significantly with V_{DS1} [Fig. 12.34(b)]. (Recall that for a high efficiency, V_{DS1} goes from near zero to $2V_{DD}$ and I_{D1} from near zero to a large value, creating considerable change in r_{O1} and C_{DB1}.) Thus, a nonlinear complex output impedance must be matched to a linear load.

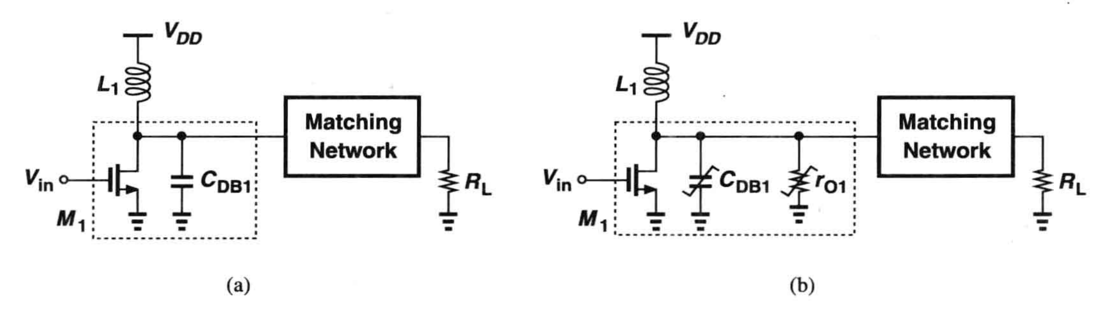


Figure 12.34 *CS stage with (a) linear drain capacitance and (b) nonlinear drain capacitance and resistance.*

Before dealing with the task of nonlinear impedance matching, let us first consider a simple case where the transistor is modeled as an ideal current source having a *linear* resistive output impedance [Fig. 12.35(a)]. For a given r_{O1}, how do we choose R_L? Let us compute the power delivered by M_1 to R_L, P_{RL}, and that consumed by the transistor's output resistance, P_{ro1}. We have

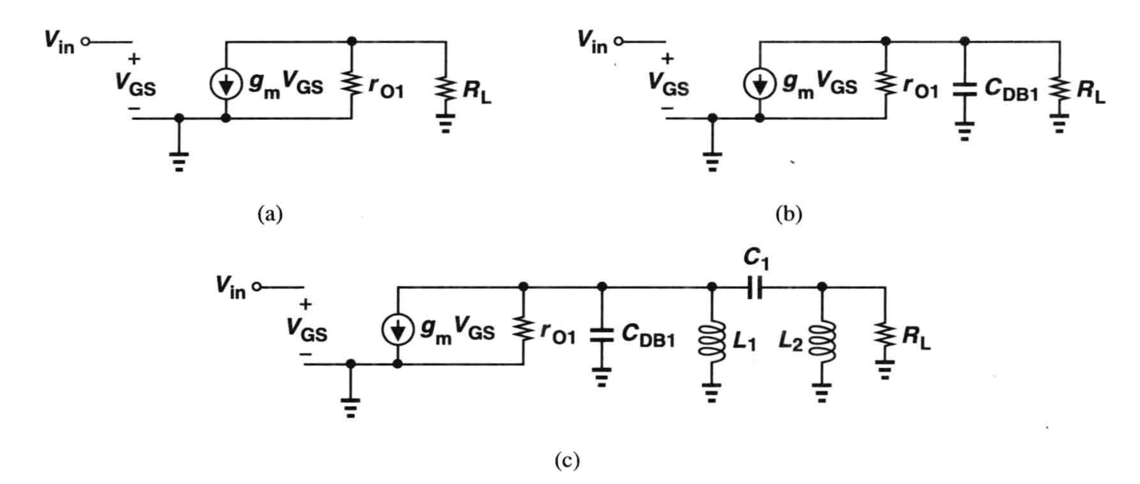


Figure 12.35 *Impedance matching with (a) simple transistor model, (b) C_{DB} included, (c) an LC network.*

$$P_{RL} = \frac{I_p^2}{2} \frac{R_L r_{O1}^2}{(R_L + r_{O1})^2}, \tag{12.60}$$

where I_p denotes the peak amplitude of the transistor's RF current. Similarly,

$$P_{ro1} = \frac{I_p^2}{2} \frac{R_L^2 r_{O1}}{(R_L + r_{O1})^2}. \tag{12.61}$$

For maximum power transfer, R_L is chosen equal to r_{O1}, yielding $P_{RL} = P_{ro1}$. That is, the transistor consumes half of the power, dropping the efficiency by a factor of two. On the other hand, since

$$\frac{P_{RL}}{P_{ro1}} = \frac{r_{O1}}{R_L}, \tag{12.62}$$

we recognize that *reducing* R_L minimizes the relative power consumed by the transistor, allowing the efficiency to approach its theoretical maximum (e.g., 50% in class A stages). The key point here is that maximum power transfer does not correspond to maximum efficiency.[5] In PA design, therefore, R_L is transformed to a value *much less* than r_{O1}.[6]

In the next step, suppose, as shown in Fig. 12.35(b), the transistor output capacitance is also included. Note that M_1 may be several *millimeters* wide for an output power level of, say, 100 mW, exhibiting large capacitances. The matching network must now provide a reactive component to cancel the effect of C_{DB1}. Figure 12.35(c) illustrates a simple example where L_1 cancels C_{DB1}, and C_1 and L_2 transform R_L to a lower value.

Now consider the general case of a *nonlinear* complex output impedance. A small-signal approximation of the impedance in the midrange of the output voltage and current can be used to obtain rough values for the matching network components, but modifying these values for maximum large-signal efficiency requires a great deal of trial and error, especially if the package parasitics must be taken into account. In practice, a more systematic approach called the "load-pull measurement" is employed.

Load-Pull Measurement Let us envision how the matching network interposed between the output transistor and the load must be designed. As conceptually shown in Fig. 12.36(a), a lossless *variable* passive network (a "tuner") can present to M_1 a complex load impedance, Z_1, whose imaginary and real parts are controlled externally. We vary Z_1 such that the power delivered to R_L remains constant and equal to P_1, thus obtaining the contour depicted in Fig. 12.36(b). A low P_1 corresponds to a broader range of $Re\{Z_1\}$ and $Im\{Z_1\}$ and hence a wider contour. Next, we seek those values of Z_1 that yield a higher output power, P_2, arriving at another (perhaps tighter) contour. These "load-pull" measurements can be repeated for increasing power levels, eventually arriving at an optimum impedance, Z_{opt}, for the maximum output power. Note that the power contours also indicate the sensitivity of P_{out} to errors in the choice of Z_1.

In the above arrangement, the input impedance of the transistor, Z_{in}, has some dependence on Z_1 due to the gate-drain capacitance of M_1. Thus, the power delivered *to* the

5. From another perspective, if power is not transferred, it is not necessarily *dissipated*.

6. Nonetheless, the PA output impedance as seen by the antenna must be somewhat close to 50 Ω to absorb reflections from the antenna. That is, PAs must typically achieve a reasonable $|S_{22}|$.

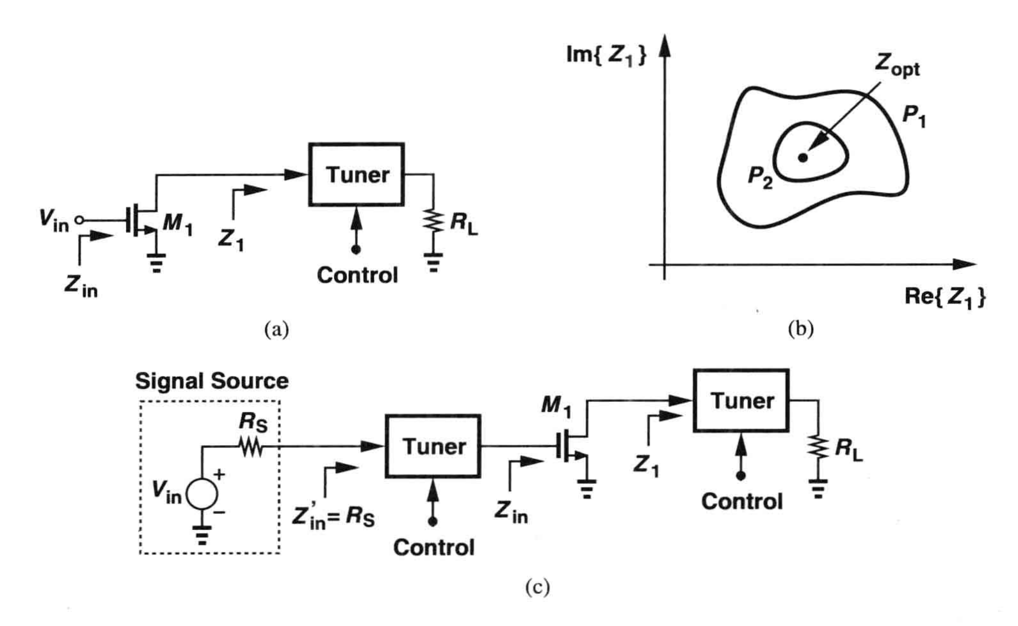


Figure 12.36 *(a) Load-pull test, (b) contours used in load-pull test, (c) computation of input and output matching impedances.*

transistor varies with Z_1, leading to a variable power gain. This effect can be avoided by inserting another tuner between the signal generator and the gate and adjusting it to obtain conjugate matching at the input for each value of Z_1 [Fig. 12.36(c)]. In a multistage PA, however, this adjustment may be unnecessary: after Z_1 reaches the optimum, Z_{in} assumes a certain value, and the preceding stage is simply designed to drive Z_{in}.

The load-pull technique has been widely used in PA design, but it requires an automated setup with precise and stable tuners. This method has three drawbacks. First, the measured results for one device size cannot be directly applied to a different size. Second, the contours and impedance levels are measured at a single frequency, failing to predict the behavior (e.g., stability) at other frequencies. Third, since the optimum choice of Z_1 in Fig. 12.36(a) does not necessarily provide peaking at higher harmonics, this technique cannot predict the efficiency and output power in the presence of harmonic termination. For these reasons, high-performance PA design using load-pull data still entails some trial and error.

12.6 BASIC LINEARIZATION TECHNIQUES

Recall from Section 12.3 that PAs designed for a high efficiency suffer from considerable nonlinearity. For relatively low output power levels, e.g., less than + 10 dBm (10 mW), we may simply back off from the PA's 1-dB compression point until the linearity reaches an acceptable value. The efficiency then falls significantly (e.g., to 10% for OFDM with 16QAM), but the absolute power drawn from the supply may still be reasonable (e.g., 100 mW). For higher output power levels, however, a low efficiency translates to a very large power consumption.

A great deal of effort has been expended on linearization techniques that offer a higher overall efficiency than back-off from the compression point does. As we will see, such techniques can be categorized under two groups: those that require *some* linearity in the PA core, and those that, in principle, can operate with arbitrarily nonlinear stages. We expect the latter to achieve a higher efficiency.

Another point observed in the following study is that linear PAs are rarely realized as negative-feedback amplifiers. This is out of concern for stability, especially if the package parasitics and their variability must be taken into account.

In this section, we present four techniques: feedforward, Cartesian feedback, predistortion, and envelope feedback. Two other techniques, namely, polar modulation and outphasing have become popular enough in modern RF design that they merit their own sections and will be studied in Sections 12.7 and 12.8, respectively.

12.6.1 Feedforward

A nonlinear PA generates an output voltage waveform that can be viewed as the sum of a linear replica of the desired signal and an "error" signal. The "feedforward" architecture computes this error and, with proper scaling, subtracts it from the output waveform [12–14]. Shown in Fig. 12.37(a) is a simple example, where the output of the main PA, V_M, is scaled by a factor of $1/A_v$, generating V_N. The input is subtracted from V_N and the result is scaled by A_v and subtracted from V_M. If $V_M = A_v V_{in} + V_D$, where V_D represents the distortion content, then

$$V_N = V_{in} + \frac{V_D}{A_v}, \tag{12.63}$$

yielding $V_p = V_D/A_v$, $V_Q = V_D$, and hence $V_{out} = A_v V_{in}$.

In practice the two amplifiers in Fig. 12.37(a) exhibit substantial phase shift at high frequencies, causing imperfect cancellation of V_D. Thus, as shown in Fig. 12.37(b), a delay stage, Δ_1, is inserted to compensate for the phase shift of the main PA, and another, Δ_2, for the phase shift of the error amplifier. The two paths leading from V_{in} to the first subtractor are sometimes called the "signal cancellation loop" and the two from M and P to the second subtractor, the "error cancellation loop."

Avoiding feedback, the feedforward topology is inherently stable if the two constituent amplifiers remain stable, the principal advantage of this architecture. Nonetheless, feedforward suffers from several shortcomings that have made its use in integrated PA design

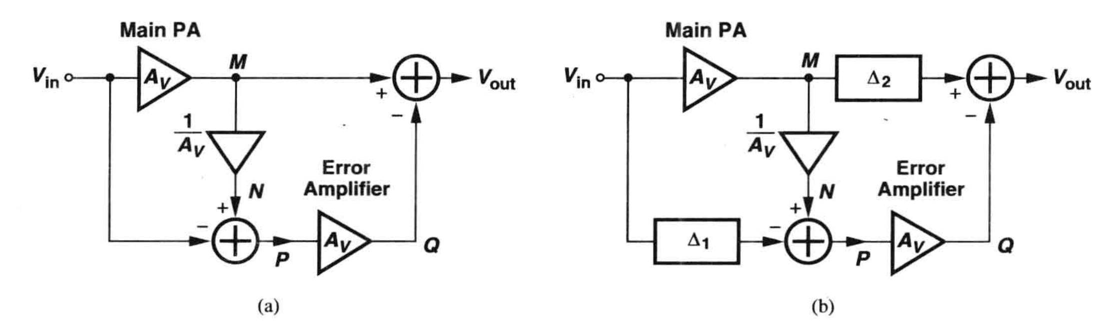


Figure 12.37 *Feedforward linearization.*

difficult. First, the analog delay elements introduce loss if they are passive or distortion if they are active, a particularly serious issue for Δ_2 as it carries a full-swing signal. Second, the loss of the output subtractor (e.g., a transformer) degrades the efficiency. For example, a loss of 1 dB lowers the efficiency by about 22%.

Example 12.16

A student surmises that the output subtraction need not introduce loss if it is performed in the current domain, e.g., as shown in Fig. 12.38. Explain the feasibility of this idea.

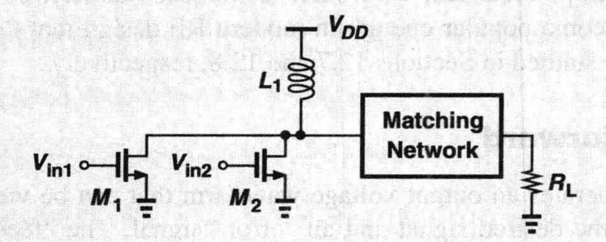


Figure 12.38 *Addition of signals in current domain.*

Solution:

Since the main PA in Fig. 12.37(b) is followed by a delay line and since performing delay in the current domain is difficult, the subtraction must inevitably occur in the voltage domain—and by means of passive devices. Thus, the idea is not practical. Other issues related to this concept are discussed later.

Third, the linearity improvement depends on the gain and phase matching of the signals sensed by each subtractor. The linearity can be measured by a two-tone test. It can be shown [12] that if the two paths from V_{in} in Fig. 12.37(b) to the inputs of the first subtractor exhibit a phase mismatch of $\Delta\phi$ and a relative gain mismatch of $\Delta A/A$, then the suppression of the magnitude of the intermodulation products in V_{out} is given by

$$E = \sqrt{1 - 2\left(1 + \frac{\Delta A}{A}\right)\cos\Delta\phi + \left(1 + \frac{\Delta A}{A}\right)^2}. \qquad (12.64)$$

For example, if $\Delta A/A = 5\%$ and $\Delta\phi = 5^\circ$, then $E = 0.102$, i.e., feedforward lowers the IM products by approximately 20 dB. The phase and gain mismatches in the error correction loop further degrade the performance.

Example 12.17

Considering the system of Fig. 12.37(b) as a "core" PA, apply another level of feedforward to further improve the linearity.

Example 12.17 (*Continued*)

Solution:

Figure 12.39 shows the "nested" feedforward architecture [15]. The core PA output is scaled by $1/A_v'$, and a delayed replica of the main input is subtracted from it. The error is scaled by A_v' and summed with the delayed replica of the core PA output.

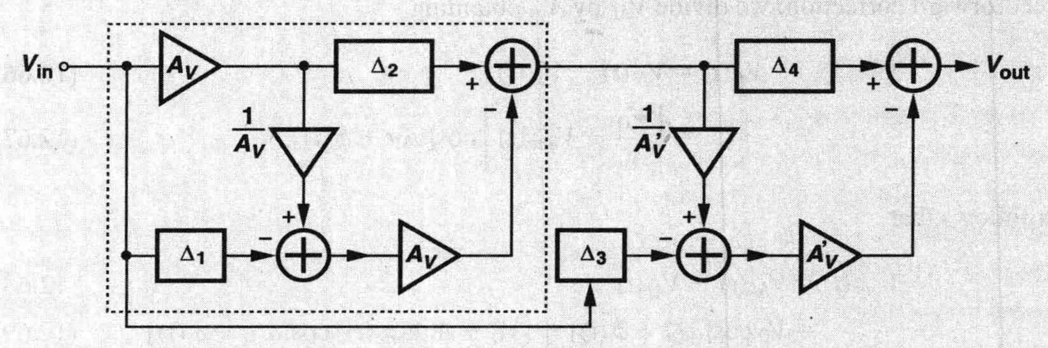


Figure 12.39 *Nested feedforward systems.*

While various calibration schemes can be conceived to deal with path mismatches, the loss of the output subtractor (and Δ_2) are the principal drawbacks of this architecture.

Example 12.18

Suppose the main PA stage in Fig. 12.37(a) is completely nonlinear, i.e., its output transistor operates as an ideal switch. Study the effect of feedforward on the PA.

Solution:

With the output transistor acting as an ideal switch, the PA removes the envelope of the signal, retaining only the phase modulation (Fig. 12.40). If $V_{in}(t) = V_{env}(t)\cos[\omega_0 t + \phi(t)]$,

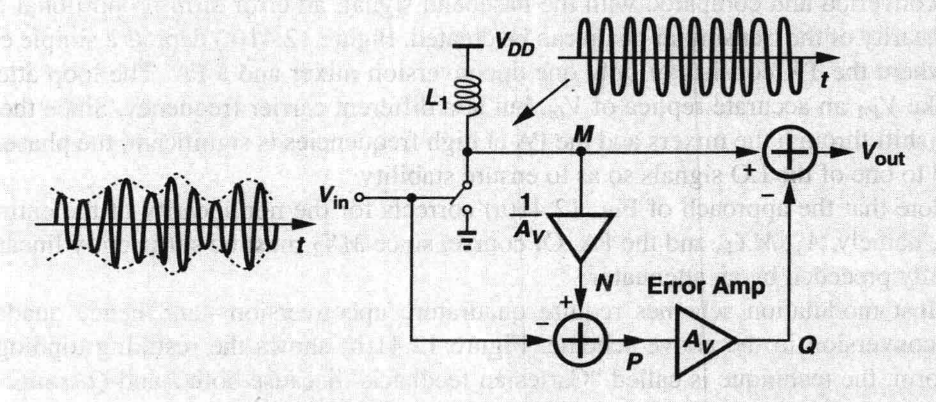


Figure 12.40 *Simplified feedforward system.*

(*Continues*)

Example 12.18 *(Continued)*

then

$$V_M(t) = V_0 \cos[\omega_0 t + \phi(t)], \quad (12.65)$$

where V_0 is constant. For such a nonlinear stage, it is difficult to define the voltage gain, A_v, because the output has little resemblance to the input. Nonetheless, let us proceed with feedforward correction: we divide V_M by A_v, obtaining

$$V_P(t) = V_N(t) - V_{in}(t) \quad (12.66)$$

$$= \left[\frac{V_0}{A_v} - V_{env}(t)\right] \cos[\omega_0 t + \phi(t)]. \quad (12.67)$$

It follows that

$$V_{out}(t) = V_M(t) - V_Q(t) \quad (12.68)$$

$$= V_0 \cos[\omega_0 t + \phi(t)] - [V_0 - A_v V_{env}(t)] \cos[\omega_0 t + \phi(t)] \quad (12.69)$$

$$= A_v V_{env}(t) \cos[\omega_0 t + \phi(t)]. \quad (12.70)$$

The output can therefore faithfully track the input with a voltage gain of A_v. Interestingly, the final output is independent of V_0.

12.6.2 Cartesian Feedback

As mentioned previously, stability issues make it difficult to apply high-frequency negative feedback around power amplifiers. However, if most of the loop gain necessary for linearization is obtained at *low* frequencies, the excess phase shift may be kept small and the system stable. In a transmitter, this is possible because the waveform processed by the PA in fact originates from upconverting a *baseband* signal. Thus, if the PA output is downconverted and compared with the baseband signal, an error term proportional to the nonlinearity of the transmitter chain can be created. Figure 12.41(a) depicts a simple example, where the TX consists of only one upconversion mixer and a PA. The loop attempts to make V_{PA} an accurate replica of V_{in}, but at a different carrier frequency. Since the total phase shift through the mixers and the PA at high frequencies is significant, the phase, θ, is added to one of the LO signals so as to ensure stability.

Note that the approach of Fig. 12.41(a) corrects for the nonlinearity of the entire TX chain, namely, A_1, MX_1, and the PA. Of course, since MX_2 must be sufficiently linear, it is typically preceded by an attenuator.

Most modulation schemes require quadrature upconversion—and hence quadrature downconversion in the above scheme. Figure 12.41(b) shows the resulting topology. In this form, the technique is called "Cartesian feedback" because both I and Q components participate in the loop.

It is instructive to compare the feedforward and Cartesian feedback topologies. The latter avoids the output subtractor and is much less sensitive to path mismatches. However,

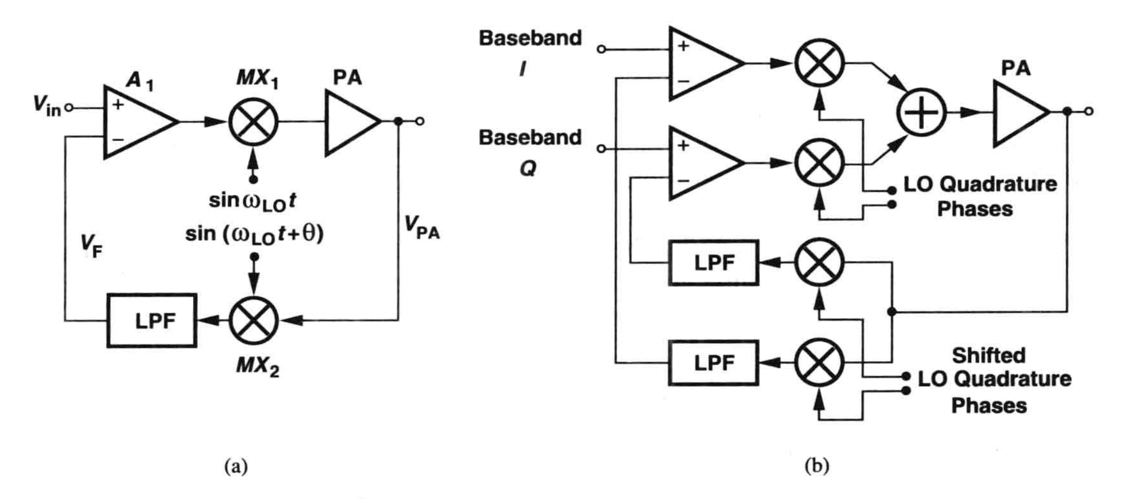


Figure 12.41 *(a) PA with translational feedback loop, (b) Cartesian feedback.*

Cartesian feedback requires *some* linearity in the PA: if a completely nonlinear PA removes the envelope, no amount of feedback can restore it.

Cartesian feedback faces a severe issue: the choice of the stabilizing LO phase shift [e.g., θ in Fig. 12.41(a)] is not straightforward because the loop phase shift varies with process and temperature. For example, while roaming toward or away from the base station, a cell phone adjusts the PA output level and, inevitably, the chip temperature, making it difficult to select a single value for θ.

12.6.3 Predistortion

If the PA nonlinear characteristics are known, it is possible to "predistort" the *input* waveform in such a manner that, after experiencing the PA nonlinearity, it resembles the ideal waveform. For example, for a PA static characteristic expressed as $y = g(x)$, predistortion subjects the input to a characteristic given by $y = g^{-1}(x)$ [Fig. 12.42(a)]. Specifically, if $g(x)$ is compressive, predistortion must expand the signal amplitude.

Predistortion suffers from three drawbacks. First, the performance degrades if the PA nonlinearity varies with process, temperature, and load impedance while the predistorter

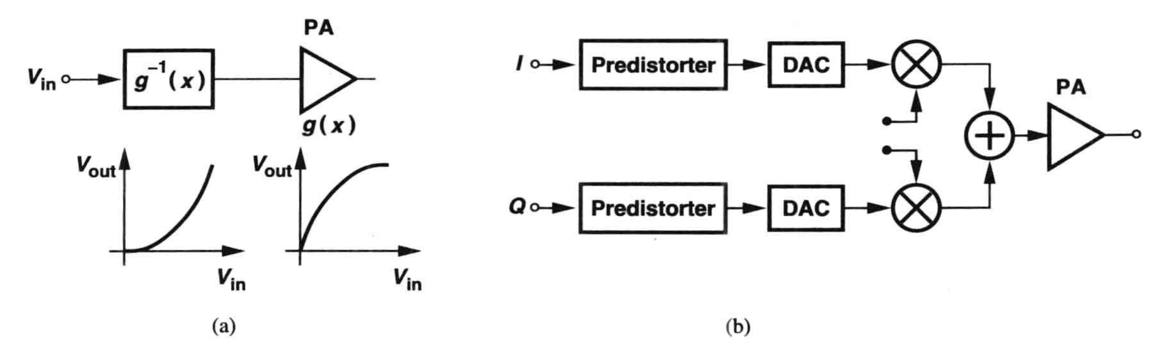


Figure 12.42 *(a) Basic predistortion concept, (b) realization in baseband.*

does not track these changes. For example, if the PA becomes more compressive, then the predistorter must become more expansive, a difficult task. Second, the PA cannot be arbitrarily nonlinear as no amount of predistortion can correct for an abrupt nonlinearity. Third, variations in the antenna impedance (e.g., how a user holds a cell phone) somewhat affect the PA nonlinearity, but predistortion provides a fixed correction.

Predistortion can also be realized in the digital domain to allow a more accurate cancellation. Illustrated in Fig. 12.42(b), the idea is to alter the baseband signal (e.g., expand its amplitude) such that it returns to its ideal waveform upon experiencing the TX chain nonlinearity. Of course, the above two issues still persist here.

Example 12.19

A student surmises that the performance of the topology shown in Fig. 12.42(a) can be improved if the predistorter is continuously informed of the PA nonlinearity, i.e., if the PA output is fed back to the predistorter. Explain the pros and cons of this idea.

Solution:

Feedback around these topologies in fact leads to architectures resembling those shown in Fig. 12.41. Depicted in Fig. 12.43 is an example, where the feedback signal produced by the low-frequency ADCs "adjusts" the predistortion.

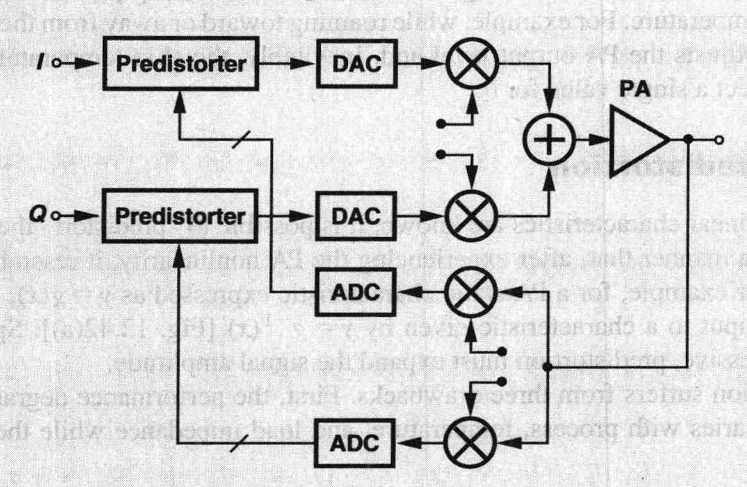


Figure 12.43 *Predistortion with feedback.*

12.6.4 Envelope Feedback

In order to reduce envelope nonlinearity (i.e., AM/AM conversion) of PAs, it is possible to apply negative feedback only to the envelope of the signal. Illustrated in Fig. 12.44, the idea is to attenuate the output by a factor of α, detect the envelope of the result, compare it with the input envelope, and adjust the gain of the signal path accordingly. With a high loop gain, the signals at A and B are nearly identical, thus forcing V_{out} to track V_{in} with a gain factor of $1/\alpha$.

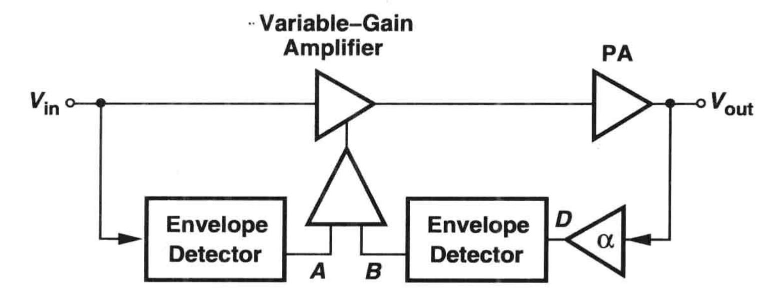


Figure 12.44 *PA with envelope feedback.*

Example 12.20

How does the distortion of the envelope detectors affect the performance of the above system?

Solution:

If the two detectors remain identical, their distortion does not affect the performance because the feedback loop still yields $V_A \approx V_B$ and hence $V_D \approx V_{in}$. This property proves greatly helpful here as typical envelope detectors suffer from nonlinearity.

Envelope Detection The reader may wonder how an envelope detector can be designed. As shown in Fig. 12.45(a), a mixer can raise the input to the power of two, yielding from

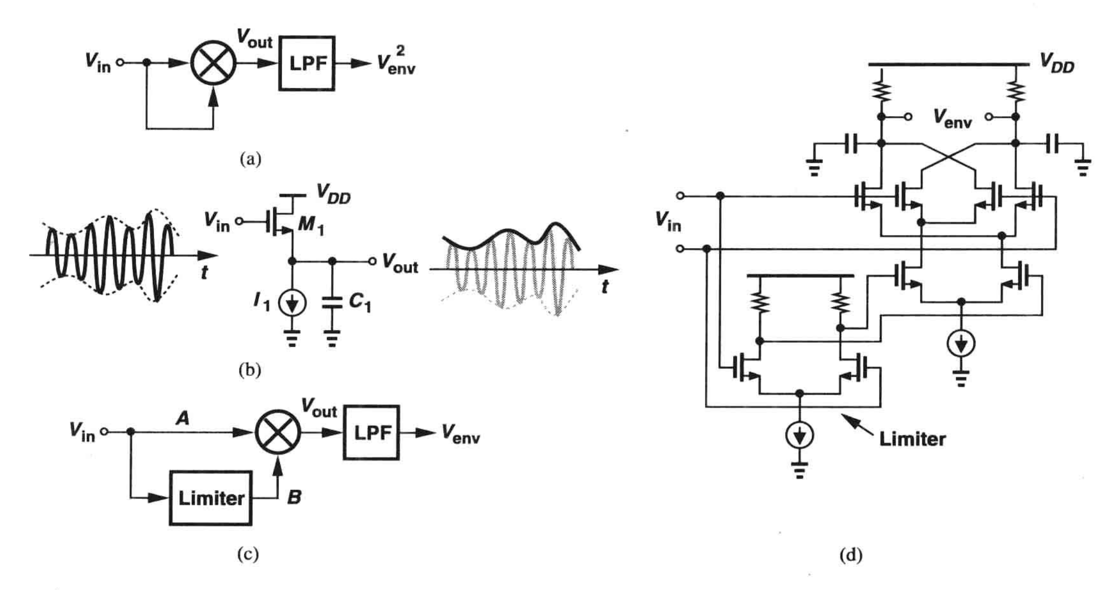


Figure 12.45 *(a) Mixer as envelope detector, (b) source follower as envelope detector, (c) limiter and mixer as envelope detector, (d) realization of (c).*

$V_{in}(t) = V_{env}(t)\cos[\omega_0 t + \phi(t)]$ the following output

$$V_{out}(t) = \beta V_{env}^2(t)\cos^2[\omega_0 t + \phi(t)] \tag{12.71}$$

$$= \beta V_{env}^2(t)\frac{1 + \cos[2\omega_0 t + 2\phi(t)]}{2}, \tag{12.72}$$

where β denotes the mixer conversion gain. Thus, the low-frequency term at the output is proportional to $V_{env}^2(t)$. Since the nonlinearity of the envelope detector in the above scheme is not critical, this topology appears a plausible choice.

Figure 12.45(b) shows an envelope detector circuit based on "peak detection." Here, the slew rate given by I_1/C_1 is chosen much much less than the carrier slew rate so that the output tracks the envelope but not the carrier. As V_{in} rises above $V_{out} + V_{TH}$, V_{out} tends to track it, but as V_{in} falls, M_1 turns off and V_{out} remains relatively constant because I_1 discharges C_1 very slowly. The dimensions of M_1 and the values of I_1 and C_1 must be chosen carefully here: if M_1 is not strong enough or C_1 is excessively large, then V_{out} fails to track the envelope itself.

A true envelope detector can be realized if the topology of Fig. 12.45(a) is modified as shown in Fig. 12.45(c). Called a "synchronous AM detector," the circuit employs a limiter in either of the signal paths, thus removing the envelope variation in that path. Denoting the signal at B by $V_0\cos[\omega_0 t + \phi(t)]$, we have

$$V_{out}(t) = \beta V_0 V_{env}(t)\cos^2[\omega_0 t + \phi(t)] \tag{12.73}$$

$$= \beta V_0 V_{env}(t)\frac{1 + \cos[2\omega_0 t + 2\phi(t)]}{2}. \tag{12.74}$$

The low-pass filter therefore produces the true envelope. Figure 12.45(d) depicts the transistor-level implementation. Here, the limiter transistors must have a small overdrive voltage so that they remove the amplitude variation. In practice, the limiter may require two or more cascaded differential pairs so as to remove envelope variations in one path leading to the mixer.

12.7 POLAR MODULATION

A linearization originally called "envelope elimination and restoration" (EER) [16] and more recently known as "polar modulation" [17] has become popular in the past ten years. This technique offers two key advantages that allow a high efficiency: (1) it can operate with an arbitrarily nonlinear output stage,[7] and (2) it does not require an output combiner (e.g., the subtractor in the feedforward topology).

12.7.1 Basic Idea

Let us begin with the original EER method. As mentioned in Chapter 3, any band-pass signal can be represented as $V_{in}(t) = V_{env}(t)\cos[\omega_0 t + \phi(t)]$, where $V_{env}(t)$ and $\phi(t)$ denote

7. It is assumed that AM/PM conversion in the output stage is negligible or can be corrected.

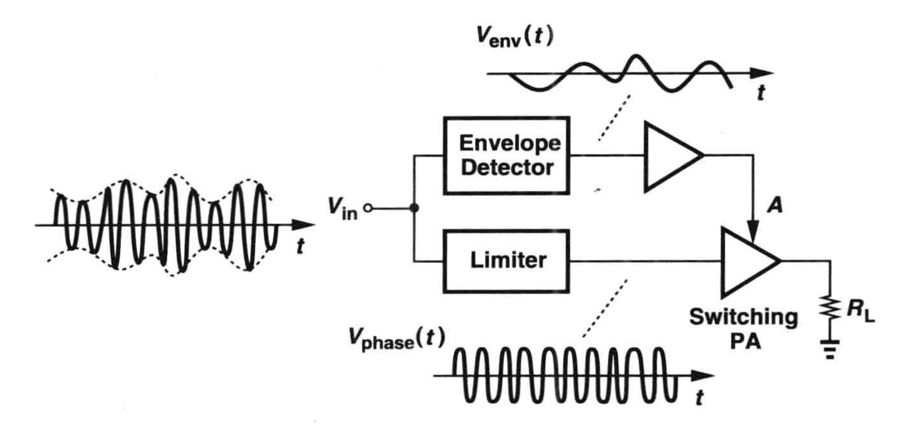


Figure 12.46 *Envelope elimination and restoration.*

the envelope and phase components, respectively. We may then postulate that we can decompose $V_{in}(t)$ into an envelope signal and a phase signal, amplify each separately, and combine the results at the end. Figure 12.46 illustrates the concept. The input signal drives both an envelope detector and a limiting stage, thus generating the envelope, $V_{env}(t)$, and the phase-modulated component, $V_{phase}(t) = V_0 \cos[\omega_0 t + \phi(t)]$. Note that the latter still contains the carrier—rather than only $\phi(t)$—even though it is called the "phase" signal. These signals are subsequently amplified and "combined" in the PA, reproducing the desired waveform. Since the output stage amplifies a constant-envelope signal, $V_{phase}(t)$, it can be nonlinear and hence efficient. This approach is also called polar modulation because it processes the signal in the form of a magnitude (envelope) component and a phase component.

How should the amplified versions of $V_{env}(t)$ and $V_{phase}(t)$ be combined in the output stage? Denoting those versions by $A_0 V_{env}(t)$ and $A_0 V_{phase}(t)$, respectively, we observe that the desired output assumes the form $A_0 V_{env}(t) \cos[\omega_0 t + \phi(t)]$, i.e., the amplitude of $A_0 V_{phase}(t)$ must be modulated by $A_0 V_{env}(t)$. It follows that the combining operation must entail multiplication or mixing rather than linear addition.

Example 12.21

A student decides that a simple mixer serves the purpose of combining and constructs the system shown in Fig. 12.47. Is this a good idea?

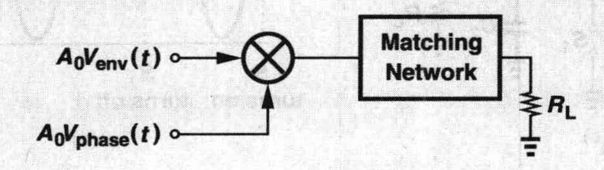


Figure 12.47 *Use of mixer to combine envelope and phase signals.*

(*Continues*)

Example 12.21 *(Continued)*

Solution:

No, it is not. Here, it is the mixer—rather than the PA core—that must deliver a high power, a very difficult task.

The combining operation is typically performed by applying the envelope signal to the *supply voltage*, V_{DD}, of the output stage—with the assumption that the output voltage swing is a function of V_{DD}. To understand this point, let us begin with the simple circuit depicted in Fig. 12.48(a), where S_1 is driven by the phase signal. When S_1 turns on, V_{out} jumps to near zero and subsequently rises exponentially toward V_{DD} [Fig. 12.48(b)]. When S_1 turns off, the instantaneous change in the inductor current yields an impulse in the output voltage. The output voltage swing is clearly a function of V_{DD}. Note the average areas under the exponential section and the impulse must be equal so that the output average remains equal to V_{DD}.

Now consider the more realistic circuit shown in Fig. 12.48(c). In this case, the output waveform somewhat resembles a sinusoid [Fig. 12.48(d)], but its amplitude is still a function of V_{DD}.

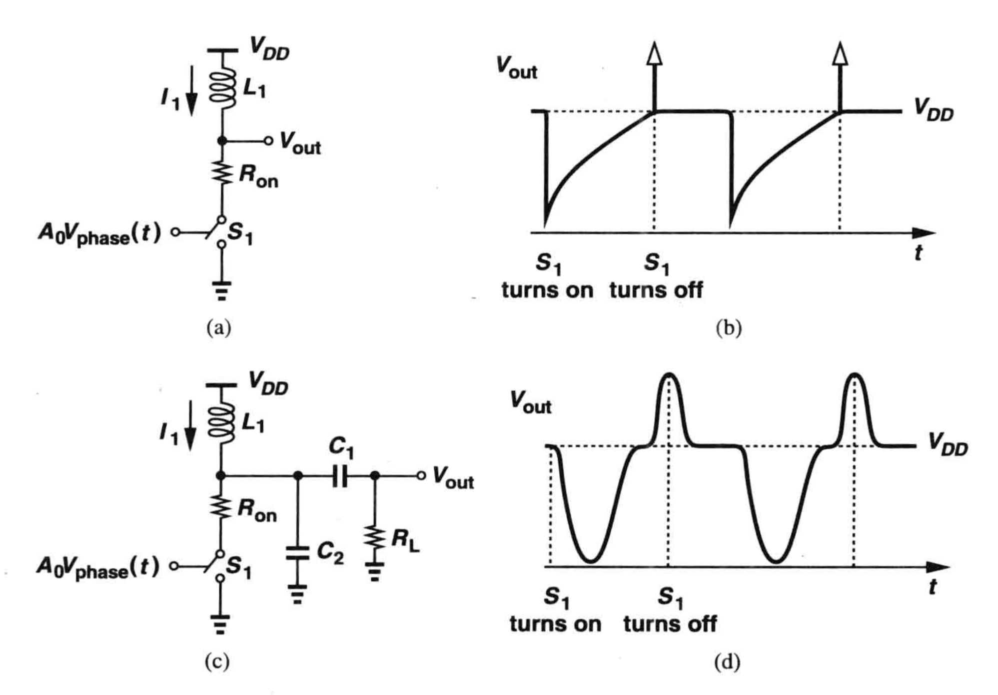


Figure 12.48 *(a) Simple model of output stage, (b) output waveform, (c) stage with capacitances and load resistance, (d) resulting output waveform.*

Example 12.22

Under what condition is the PA output swing not a function of V_{DD}?

Solution:

If the output transistor acts as a voltage-dependent *current source* (e.g., a MOSFET operating in saturation), then the output swing is only a weak function of V_{DD}. In other words, all PA classes that employ the output transistor as a current source fall in this category and are not suited to EER.

The foregoing observations lead to the conceptual combining circuit shown in Fig. 12.49(a), where the envelope signal directly drives the supply node of the PA stage. The large current flowing through this stage requires a buffer in this path, but efficiency considerations demand minimal voltage headroom consumption by the buffer. As an example, the arrangement in Fig. 12.49(b) incorporates a voltage-dependent resistor, M_2, to modulate $V_{DD,PA}$, in proportion to $A_0V_{env}(t)$. For an average current of I_0 through L_1 and an average voltage drop of V_0 across the drain-source resistance of M_2, this device dissipates a power of I_0V_0, lowering the efficiency. Thus, M_2 is typically a very wide transistor.

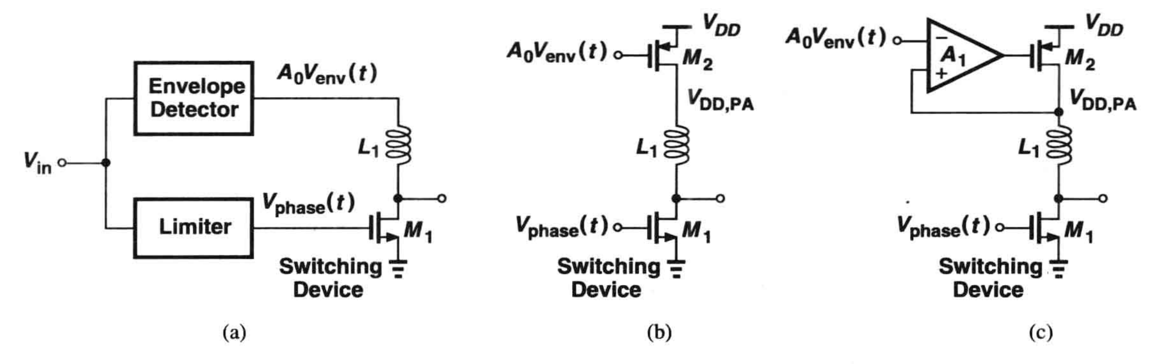


Figure 12.49 *(a) Partial realization of EER, (b) output stage with envelope-controlled load, (c) local envelope feedback.*

Does the circuit of Fig. 12.49(b) guarantee that $V_{DD,PA}$ tracks $A_0V_{env}(t)$ faithfully? No, it does not: in this "open-loop" control, $V_{DD,PA}$ is a function of various device parameters. This issue becomes more serious if the PA must provide a variable output level because changing the current of the output stage also alters $V_{DD,PA}$. We may modify the stage to the "closed-loop" control shown in Fig. 12.49(c), where amplifier A_1 introduces a high loop gain so that $V_{DD,PA} \approx A_0V_{env}(t)$. Of course, A_1 must accommodate an input common-mode level near V_{DD}.

12.7.2 Polar Modulation Issues

Polar modulation entails a number of issues. First, the mismatch between the delays of the envelope and phase paths corrupts the signal in Fig. 12.46. To formulate this effect, we

assume a delay mismatch of ΔT and express the output as

$$V_{out}(t) = A_0 V_{env}(t - \Delta T)\cos[\omega_0 t + \phi(t)]. \tag{12.75}$$

For a small ΔT, $V_{env}(t - \Delta T)$ can be approximated by the first two terms in its Taylor series:

$$V_{env}(t - \Delta T) \approx V_{env}(t) - \Delta T \frac{dV_{env}(t)}{dt}. \tag{12.76}$$

It follows that

$$V_{out}(t) \approx A_0 V_{env}(t)\cos[\omega_0 t + \phi(t)] - \Delta T \frac{dV_{env}(t)}{dt}\cos[\omega_0 t + \phi(t)]. \tag{12.77}$$

The corruption is therefore proportional to the derivative of the envelope signal, leading to substantial spectral regrowth because the spectrum of $V_{env}(t)$ is equivalently multiplied by ω^2. For example, in an EDGE system, a delay mismatch of 40 ns allows only 5 dB of margin between the output spectrum and the required spectral mask [18].

The problem of delay mismatch is a serious one because the two paths in Fig. 12.46 employ different types of circuits operating at vastly different frequencies: the envelope path contains an envelope detector and a low-frequency buffer, whereas the phase path includes a limiter and an output stage.

The second issue relates to the linearity of the envelope detector. Unlike the feedback topology of Fig. 12.44, the polar TX in Fig. 12.46 relies on precise reconstruction of $V_{env}(t)$ by the envelope detector. As shown in Problem 12.6, this circuit's nonlinearity produces spectral regrowth.

The third issue concerns the operation of limiters at high frequencies. In general, a nonlinear circuit having a finite bandwidth introduces AM/PM conversion, i.e., exhibits a phase shift that depends on the input amplitude. For example, consider the differential pair shown in Fig. 12.50(a), where the bandwidth is defined by the output pole, $\omega_p = 1/(R_1 C_1)$. If the input is a small sinusoidal signal at ω_0, then the differential output current is also a sinusoid, experiencing a phase shift of

$$|\theta_1| = \tan^{-1}(R_1 C_1 \omega_0) \tag{12.78}$$

as it is converted to voltage. For $\omega_0 \ll \omega_p$,

$$|\theta_1| \approx R_1 C_1 \omega_0. \tag{12.79}$$

Now, if the circuit senses a large input sinusoid [Fig. 12.50(b)] such that M_1 and M_2 produce nearly rectangular drain current waveforms, then the *delay* between the input and output is approximately equal to[8]

$$\Delta T = R_1 C_1 \ln 2. \tag{12.80}$$

8. We define the delay as that between the times at which the input and the output reach 50% of their full swings.

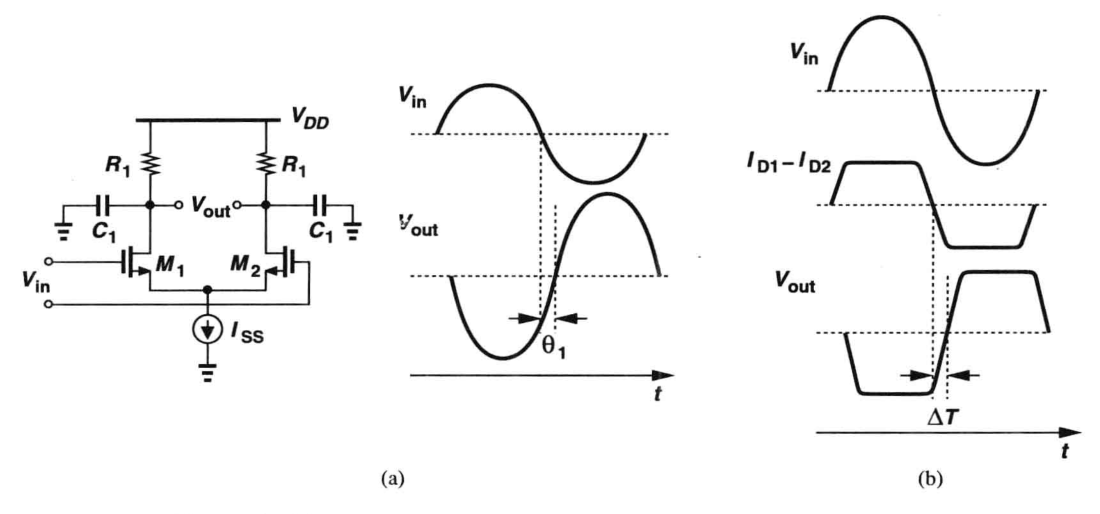


Figure 12.50 *Limiting stage with (a) small and (b) large input swings.*

Expressing this result in radians, we have

$$|\theta_2| = R_1 C_1 \omega_0 \ln 2. \tag{12.81}$$

Comparison of Eqs. (12.79) and (12.81) reveals that the phase shift decreases as the input amplitude increases. Thus, the limiter in Fig. 12.46 may corrupt the phase signal by the large excursions in the envelope.

The fourth issue stems from the variation of the output node capacitance (C_{DB}) in Fig. 12.49(c) by the *envelope* signal. As $V_{DD,PA}$ swings up and down to track $A_0 V_{env}(t)$, C_{DB} varies and so does the phase shift from the gate of M_1 to its drain, ϕ_0 (Fig. 12.51). That is, the phase signal is corrupted by the envelope signal. This effect can be quantified as follows. We recognize that the variation of C_{DB} alters the resonance frequency, ω_1, at the output node. We can therefore express the dependence of ϕ_0 upon the drain voltage as a straight line having a slope of[9]

$$\frac{d\phi_0}{dV_X} = \frac{dC_{DB}}{dV_X} \cdot \frac{d\omega}{dC_{DB}} \cdot \frac{d\phi_0}{d\omega}. \tag{12.82}$$

Figure 12.51 *AM/PM conversion due to output capacitance nonlinearity.*

9. This is equivalent to approximating $\phi_0(V_X)$ by the first two terms of its Taylor expansion.

The first derivative on the right-hand side can readily be found, e.g., from

$$C_{DB} = \frac{C_{DB0}}{\left(1 + \frac{V_X}{V_B}\right)^m}, \tag{12.83}$$

where V_B denotes the junction built-in potential and m is typically around 0.4. The second derivative, $d\omega/dC_{DB}$, is obtained from $\omega_1 = 1/\sqrt{L_1 C_{DB}}$ as

$$\frac{d\omega}{dC_{DB}} = \frac{-1}{2\sqrt{L_1 C_{DB}}} \cdot \frac{1}{L_1 C_{DB}} \tag{12.84}$$

$$= -\frac{1}{2}\omega_1^3. \tag{12.85}$$

Finally, $d\phi_0/d\omega$ is computed from the quality factor, Q, of the output network (Chapter 8); that is,

$$Q = \frac{\omega_1}{2}\frac{d\phi_0}{d\omega}, \tag{12.86}$$

and hence

$$\frac{d\phi_0}{d\omega} = \frac{2Q}{\omega_1}. \tag{12.87}$$

It follows that

$$\frac{d\phi_0}{dV_X} = -Q\omega_1^2 \frac{dC_{DB}}{dV_X}. \tag{12.88}$$

To the first order,

$$\phi_0(t) = A_0 V_{env}(t)\frac{d\phi_0}{dV_X} + \cdots. \tag{12.89}$$

As mentioned earlier, another issue in polar modulation is the efficiency (and voltage headroom) reduction due to the envelope buffer [M_2 in Fig. 12.49(c)]. We will see below that, among the issues outlined above, only the last one defies design techniques and becomes the bottleneck at low supply voltages.

12.7.3 Improved Polar Modulation

The advent of RF IC technology has also improved polar transmitters considerably. In this section, we study a number of techniques that address the issues described in the previous section. The key principle here is to expand the design horizon to include the entire transmitter chain rather than merely the RF power amplifier.

In the conceptual approach depicted in Fig. 12.46, we attempted to decompose the RF signal into envelope and phase components, thus facing limiter's AM/PM conversion. Let us instead perform this decomposition in the *baseband*. For an RF waveform $V_{env}(t)\cos[\omega_0 t + \phi(t)]$, the quadrature baseband signals are given by

$$x_{BB,I}(t) = V_{env}(t)\cos[\phi(t)] \tag{12.90}$$

$$x_{BB,Q}(t) = V_{env}(t)\sin[\phi(t)]. \tag{12.91}$$

Thus,

$$V_{env}(t) = \sqrt{x_{BB,I}^2(t) + x_{BB,Q}^2(t)} \tag{12.92}$$

$$\phi(t) = \tan^{-1}\frac{x_{BB,Q}(t)}{x_{BB,I}(t)}. \tag{12.93}$$

In other words, the digital baseband processor can generate $V_{env}(t)$ and $\phi(t)$ either directly or from the I and Q components, obviating the need for decomposition in the RF domain.

While $V_{env}(t)$ can now be applied to modulate the PA power supply, $\phi(t)$ does not easily lend itself to upconversion to radio frequencies. The following example illustrates this point.

Example 12.23

In our study of frequency-modulated or phase-modulated transmitters in Chapter 3, we encountered two architectures, namely, direct VCO modulation and quadrature upconversion. Can these architectures be utilized in a polar modulation system?

Solution:

First, consider applying the phase information to the control line of a VCO. The integration performed by the VCO requires that $\phi(t)$ be first differentiated [Fig. 12.52(a)]. We have

$$V_{phase}(t) = V_0 \cos\left(\omega_0 t + K_{VCO}\int \frac{1}{K_{VCO}}\frac{d\phi}{dt}dt\right) \tag{12.94}$$

$$= V_0 \cos[\omega_0 t + \phi(t)]. \tag{12.95}$$

However, as explained in Chapter 3, since both the full-scale swing of $d\phi/dt$ (in the analog domain) and K_{VCO} are poorly-defined, so is the bandwidth of $V_{phase}(t)$. Also, the free-running operation of the VCO during modulation may shift the carrier frequency from its desired value.

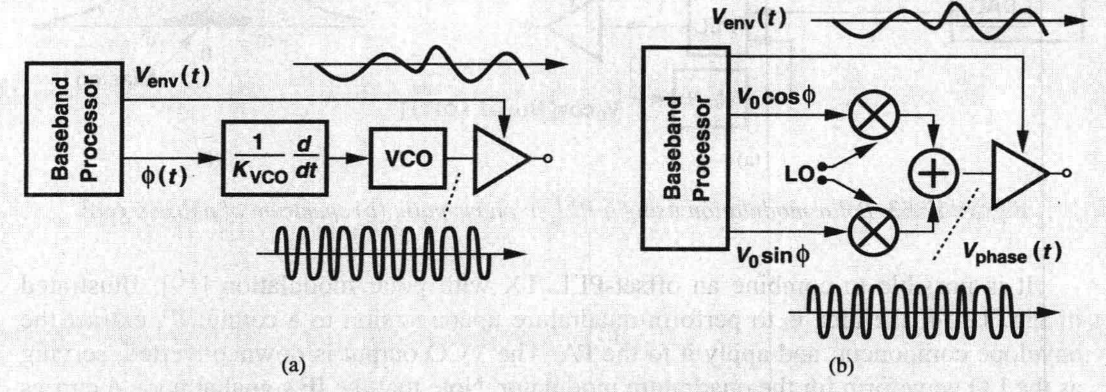


Figure 12.52 *Polar modulation using baseband signal separation and (a) a VCO, or (b) a quadrature upconverter.*

(Continues)

Example 12.23 *(Continued)*

Now, consider a quadrature modulator, as stipulated in Chapter 3 for GMSK. In this case, $V_{phase}(t)$ is expressed as

$$V_{phase}(t) = V_0 \cos \omega_0 t \cos \phi - V_0 \sin \omega_0 t \sin \phi; \tag{12.96}$$

i.e., so that $V_0 \cos \phi$ and $V_0 \sin \phi$ are produced by the baseband and upconverted by quadrature mixers [Fig. 12.52(b)]. However, as mentioned in Chapter 4, this approach may still introduce significant noise in the receive band because the noise of the mixers is upconverted and amplified by the PA.

In addition to direct VCO modulation and quadrature upconversion, we studied in Chapter 9 a number of techniques leading to the offset-PLL TX. For example, we contemplated a PLL as a means of upconversion of the phase signal. Figure 12.53(a) depicts an architecture combining that idea with polar modulation. In this case, the phase signal produced by the baseband processor is located at a finite carrier frequency, ω_{IF}, and its phase excursion is scaled down by a factor of N. The PLL thus generates an output given by

$$V_{PLL}(t) = V_0 \cos[N\omega_{IF}t + \phi(t)], \tag{12.97}$$

where $N\omega_{IF}$ is chosen equal to the desired carrier frequency. The value of ω_{IF} must remain between two bounds: (1) it must be low enough to avoid imposing severe speed-power trade-offs on the baseband DAC, and (2) it must be high enough to avoid aliasing [Fig. 12.53(b)].

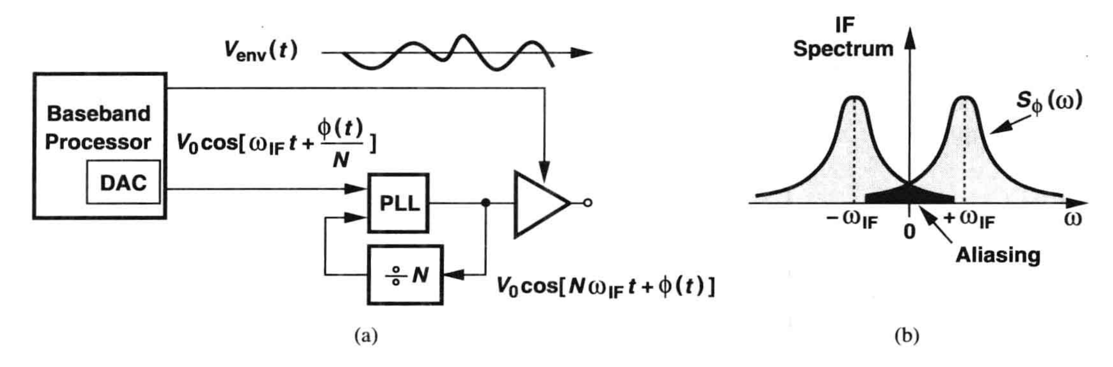


Figure 12.53 *Polar modulation using a PLL in phase path, (b) spectrum of phase signal.*

It is possible to combine an offset-PLL TX with polar modulation [19]. Illustrated in Fig. 12.54, the idea is to perform quadrature upconversion to a certain IF, extract the envelope component, and apply it to the PA. The VCO output is downconverted, serving as the LO waveform for the quadrature modulator. Note that the IF signal at node A carries little phase modulation because the PLL feedback forces the phase at A to track that of f_{REF} (an unmodulated reference). With proper choice of the PLL bandwidth, the output noise in the receive band is determined primarily by the VCO design.

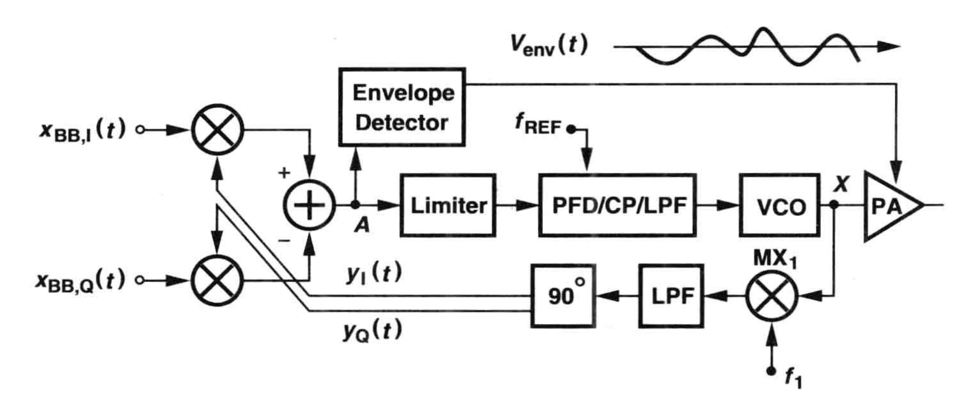


Figure 12.54 *Polar modulation with phase feedback.*

Example 12.24

How can the architecture of Fig. 12.54 be modified so as to avoid an envelope detector?

Solution:

If the quadrature upconverter senses only the baseband phase information [as in Fig. 12.52(b)], then the envelope can also come from the baseband. Figure 12.55 shows such an arrangement, where the envelope component is directly produced by the baseband processor.

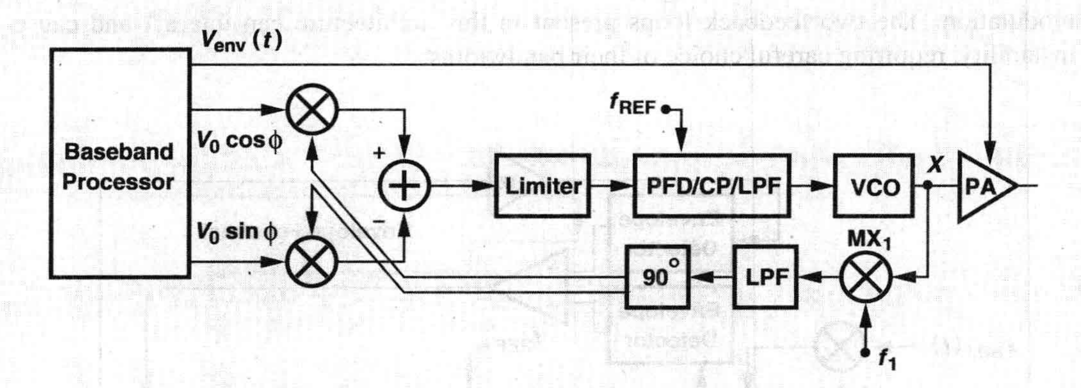


Figure 12.55 *Polar modulation without envelope detection.*

The polar modulation architectures studied above still fail to address two issues, namely, poor definition of the PA output envelope and the corruption due to the PA's AM/PM conversion (e.g., due to the output capacitance nonlinearity). We must therefore apply feedback to sense and correct these effects. As shown in Fig. 12.49(c), the envelope can be controlled precisely by means of a feedback buffer driving the supply rail of the PA. Alternatively, as in the envelope feedback architecture of Fig. 12.44, the output envelope

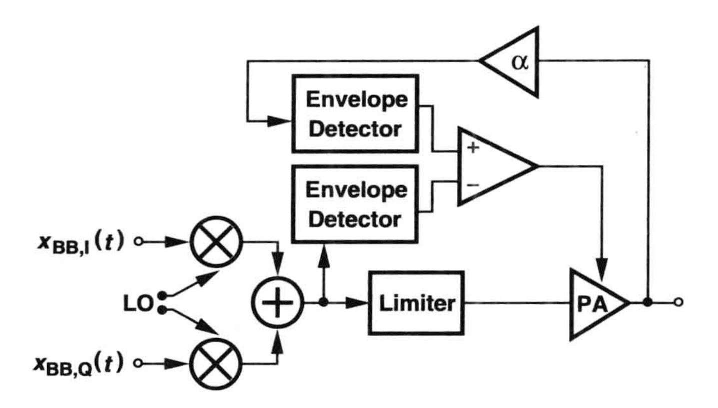


Figure 12.56 *Polar modulation with envelope feedback.*

can be compared with the input envelope. Figure 12.56 depicts the resulting arrangement. The PA output voltage swing is scaled by a factor of α, applied to an envelope detector, and compared with the IF envelope. The feedback loop thus forces a faithful (scaled) replica of the IF envelope at the PA output. The envelope detectors can be realized as shown in Figs. 12.45(c) and (d).

In order to correct the PA's AM/PM conversion, the PA output phase must appear *within* the PLL, i.e., the PLL feedback path must sense the PA output rather than the VCO output. Illustrated in Fig. 12.57, such an architecture impresses the baseband phase excursions on the PA output by virtue of the high loop gain of the PLL. In other words, if the PA introduces AM/PM conversion, the PLL still guarantees that the phase at X tracks the baseband phase modulation. The two feedback loops present in this architecture can interact and cause instability, requiring careful choice of their bandwidths.

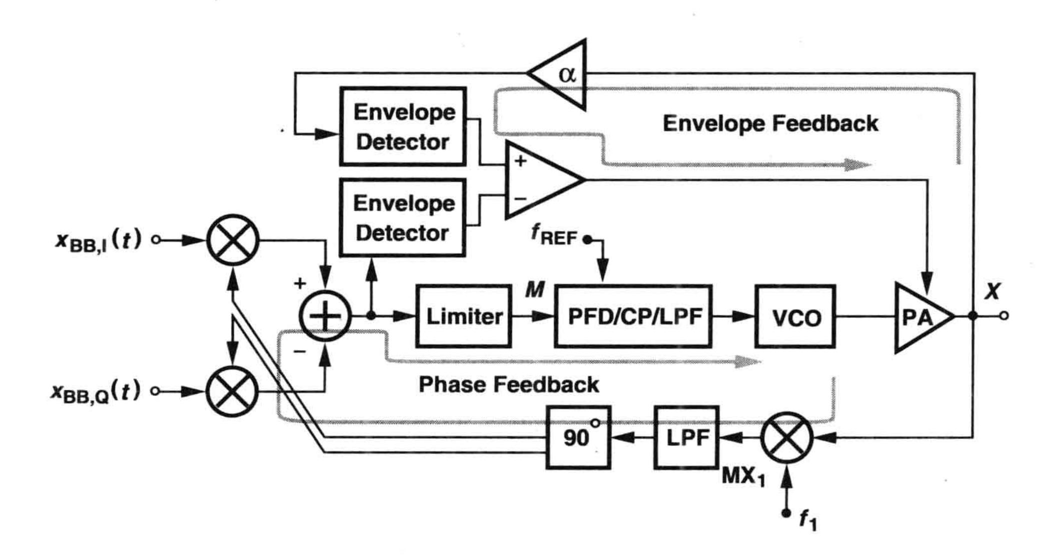


Figure 12.57 *Polar modulation with phase and envelope feedback.*

Example 12.25

Identify the drawbacks of the architecture shown in Fig. 12.57.

Solution:

A critical issue here relates to the need for power control. Since the PA output level must be variable (by about 30 dB in GSM/EDGE and 60 dB in CDMA), the swing applied to mixer MX_1 may prove insufficient at the lower end of the power range, degrading the stability of the loop. For example, for a maximum peak-to-peak swing of 2 V at X and 30 dB of power range, the minimum swing sensed by MX_1 is about 66 mV_{pp}. To resolve this issue, a limiter must be interposed between the PA and MX_1, but we recall from Fig. 12.50 that limiters introduce considerable AM/PM conversion if their input senses a wide range of amplitudes. Of course, the limiter's AM/PM conversion is not corrected by the loop.

Another drawback of the architecture is that the independent envelope and phase loops may exhibit substantially different delays, exacerbating the delay mismatch effect formulated by Eq. (12.77). In other words, the delay through the envelope detector, the error amplifier, and the supply modulation device in Fig. 12.57 may be arbitrarily different from that through the limiter, with no correction provided by the two loops.

Other Issues The architecture of Fig. 12.57 or its variants [19] resolve some of the polar modulation issues identified in Section 12.7.2. However, several other challenges remain that merit attention.

First, the bandwidths of the envelope and phase signal paths must be chosen carefully. The key point here is that each of these components occupies a *larger* bandwidth than the overall composite modulated signal. As an example, Fig. 12.58 plots the spectra of the individual components and the composite signal along with the spectral mask for an EDGE system [18]. We note that the envelope spectrum exceeds the mask in a few regions and, more importantly, the phase spectrum consumes a much broader bandwidth. If the envelope and phase paths do not provide sufficient bandwidth, then the two components are not combined properly and the final PA output suffers from spectral regrowth, possibly

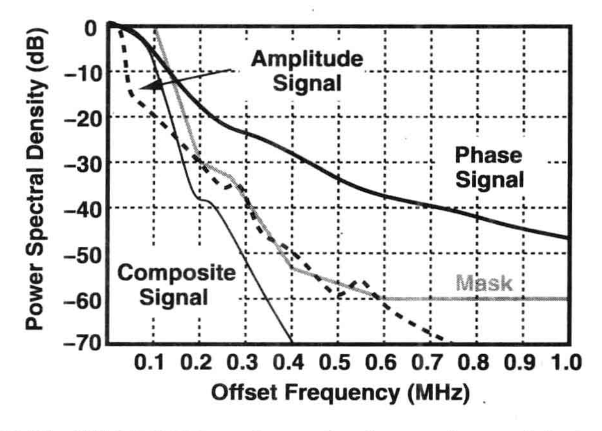


Figure 12.58 *GSM/EDGE mask margins for a polar modulation system.*

violating the spectral mask. For example, if in an EDGE system the AM and PM path bandwidths are equal to 1 MHz and 3 MHz, respectively, then the output spectrum bears only a 2-dB margin with respect to the mask [18].

While the foregoing considerations call for a large bandwidth in the two paths, we must recall that the PLL specifically serves to reduce the noise in the receive band and, therefore, cannot have a large bandwidth. The trade-off between spectral regrowth and noise in the RX band in turn dictates tight control over the PLL bandwidth. Since the dependence of the charge pump current and K_{VCO} upon process and temperature leads to significant bandwidth variation, some means of bandwidth calibration is often necessary [18].

The second issue relates to the leakage of the PM signal to the output as an *additive* component. For example, suppose, as shown in Fig. 12.59, the VCO inductor couples a fraction of the PM signal to an inductor (or a pad) at the output of the PA [18].

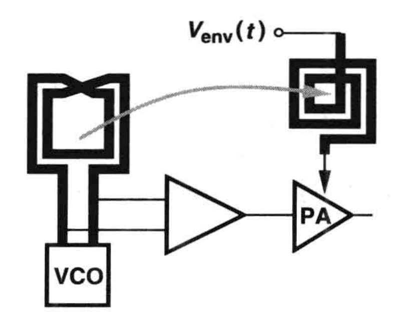


Figure 12.59 *Phase signal leakage path.*

Noting the broad bandwidth of the phase signal in Fig. 12.58, we recognize that this leakage produces considerable spectral regrowth if it does not experience proper envelope modulation [18]. This phenomenon can be readily formulated as

$$V_{out}(t) = AV_{env}(t)\cos[\omega_0(t) + \phi(t)] + V_1\cos[\omega_0 t + \phi(t)], \tag{12.98}$$

where the second term represents the additive leakage.

The third issue concerns dc offsets in the envelope path [18]. If the envelope produced by the envelope detector has an offset, V_{OS}, then the PA output is given by

$$V_{out}(t) = A_0[V_{env}(t) + V_{OS}]\cos[\omega_0 t + \phi(t)]. \tag{12.99}$$

That is, the output contains a PM leakage component equal to $A_0 V_{OS}\cos[\omega_0 t + \phi(t)]$, which must be minimized so as to avoid spectral regrowth. For example, in an EDGE system, V_{OS} must remain below 0.2% of the peak of $V_{env}(t)$ to allow sufficient margin for other errors [18]. Of course, if the output power must be variable, such a condition must hold even for the lowest output level, a difficult task.

12.8 OUTPHASING

12.8.1 Basic Idea

It is possible to avoid envelope variations in a PA by decomposing a variable-envelope signal into two *constant*-envelope waveforms. Called "outphasing" in [20] and "linear

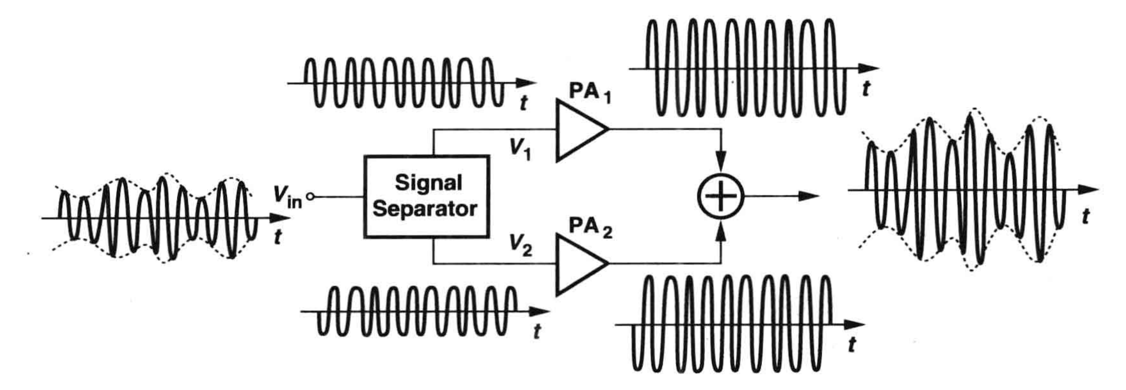


Figure 12.60 *Basic outphasing.*

amplification with nonlinear components" (LINC) in [21], the idea is that a band-pass signal $V_{in}(t) = V_{env}(t)\cos[\omega_0 t + \phi(t)]$ can be expressed as the sum of two phase-modulated components (Fig. 12.60),

$$V_{in}(t) = V_{env}(t)\cos[\omega_0 t + \phi(t)] \quad (12.100)$$

$$= V_1(t) + V_2(t), \quad (12.101)$$

where

$$V_1(t) = \frac{V_0}{2}\sin[\omega_0 t + \phi(t) + \theta(t)] \quad (12.102)$$

$$V_2(t) = -\frac{V_0}{2}\sin[\omega_0 t + \phi(t) - \theta(t)], \quad (12.103)$$

and

$$\theta(t) = \sin^{-1}\frac{V_{env}(t)}{V_0}. \quad (12.104)$$

Thus, if $V_1(t)$ and $V_2(t)$ are generated from $V_{in}(t)$, amplified by means of nonlinear stages, and subsequently added, the output contains the same envelope and phase information as does $V_{in}(t)$.

Generation of $V_1(t)$ and $V_2(t)$ from $V_{in}(t)$ requires substantial complexity, primarily because their phase must be modulated by $\theta(t)$, which itself is a nonlinear function of $V_{env}(t)$. The use of nonlinear frequency-translating feedback loops has been proposed [21, 22], but loop stability issues limit the feasibility of these techniques. A more practical approach [23] considers $V_1(t)$ and $V_2(t)$ as

$$V_1(t) = V_I(t)\cos[\omega_0 t + \phi(t)] + V_Q(t)\sin[\omega_0 t + \phi(t)] \quad (12.105)$$

$$V_2(t) = -V_I(t)\cos[\omega_0 t + \phi(t)] + V_Q(t)\sin[\omega_0 t + \phi(t)], \quad (12.106)$$

where the baseband components are given by

$$V_I(t) = \frac{V_{env}(t)}{2} \tag{12.107}$$

$$V_Q(t) = \sqrt{V_0^2 - \frac{V_{env}^2(t)}{2}}. \tag{12.108}$$

Since the nonlinear operation required to produce $V_Q(t)$ can be performed in the baseband (e.g., using a look-up ROM), this method can simply employ quadrature upconversion to generate $V_1(t)$ and $V_2(t)$.

Example 12.26

Construct a complete outphasing transmitter.

Solution:

From our study of GMSK modulation techniques in Chapter 3, we recall that the phase component, $\phi(t)$, should also be realized in the baseband rather than impressed on the LO. We therefore expand the original equations, (12.102) and (12.103), respectively, as follows

$$V_1(t) = \frac{V_0}{2}\cos[\phi(t) + \theta(t)]\sin\omega_0 t + \frac{V_0}{2}\sin[\phi(t) + \theta(t)]\cos\omega_0 t \tag{12.109}$$

$$V_2(t) = -\frac{V_0}{2}\cos[\phi(t) - \theta(t)]\sin\omega_0 t - \frac{V_0}{2}\sin[\phi(t) - \theta(t)]\cos\omega_0 t. \tag{12.110}$$

The TX is thus constructed as shown in Fig. 12.61.

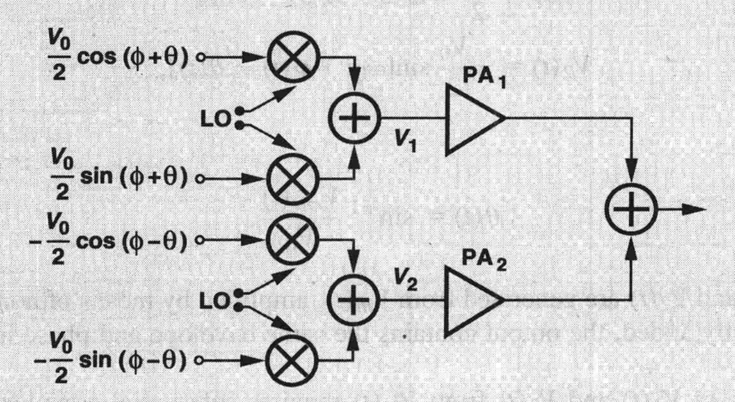


Figure 12.61 *Outphasing transmitter.*

The outphasing architecture can operate with completely nonlinear PA stages, an important attribute similar to that of polar modulation. A critical advantage of outphasing is that it does not require supply modulation, saving the efficiency and headroom lost in the envelope buffer necessary in polar modulation. Unfortunately, the summation of the outputs in the outphasing technique entails power loss (as in the feedforward topology).

12.8.2 Outphasing Issues

In addition to the output summation problem, outphasing must deal with a number of other issues. First, the gain and phase mismatches between the two paths in Fig. 12.60 result in spectral regrowth at the output. Representing the two mismatches by ΔV and $\Delta\theta$, respectively, we have

$$V_1(t) = \left(\frac{V_0}{2} + \Delta V\right)\sin[\omega_0 t + \phi(t) + \theta(t) + \Delta\theta] \tag{12.111}$$

$$V_2(t) = \frac{V_0}{2}\sin[\omega_0 t + \phi(t) - \theta(t)]. \tag{12.112}$$

If $\Delta\theta \ll 1$ radian, then

$$V_1(t)+V_2(t) = V_{env}(t)\cos[\omega_0 t + \phi(t)] + \Delta V\sin[\omega_0 t+\phi(t) + \theta(t)] - \Delta\theta\frac{V_0}{2}\cos[\omega_0 t + \phi(t)+\theta(t)]. \tag{12.113}$$

The last two terms on the right-hand side create spectral growth because they exhibit a much larger bandwidth than the composite signal (the first term).

Example 12.27

Identify the sources of mismatch in the architecture of Fig. 12.61.

Solution:

To avoid LO mismatch, the two quadrature upconverters must share the LO phases. The remaining sources include the mixers, the PAs, and the output summing mechanism.

The second issue concerns the required bandwidth of each path in Fig. 12.60. Since $V_1(t)$ and $V_2(t)$ experience large phase excursions, $\phi(t) \pm \theta(t)$ (when ϕ and θ "beat"), these two signals occupy a large bandwidth. Recall from the EDGE spectra in Fig. 12.58 that the bandwidth of a component of the form $\cos[\omega_0 t + \phi(t)]$ is several times that of the composite signal. This is exacerbated in outphasing by the additional phase, $\theta(t)$.

Example 12.28

A student attempts to reduce the excursions of $\theta(t)$ by selecting a scaling voltage of $V_a > V_0$ in Eq. (12.104):

$$\theta(t) = \sin^{-1}\frac{V_{env}(t)}{V_a}. \tag{12.114}$$

Explain the effect on the overall TX. Assume the baseband waveforms are generated according to (12.109) and (12.110), i.e., with an amplitude of $V_0/2$.

(Continues)

Example 12.28 (Continued)

Solution:

If $\theta(t)$ is scaled down while the amplitude of the baseband signals remains constant, the composite output amplitude falls. In Problem 12.9, we show that Eq. (12.113) must now be written as

$$V_1(t) + V_2(t) = \frac{V_0}{V_a} V_{env}(t)\cos[\omega_0 t + \phi(t)] + \Delta V \sin[\omega_0 t + \phi(t) + \theta(t)] - \Delta\theta \frac{V_0}{2}\cos[\omega_0 t + \phi(t) + \theta(t)]. \tag{12.115}$$

It follows that the effect of mismatches becomes more pronounced as V_a increases and $\theta(t)$ is scaled down.

The third issue relates to the interaction between the two PAs through the output summing device. The signal traveling through one PA may affect that through the other, resulting in spectral regrowth and even corruption. To understand this point, let us consider the simple summation shown in Fig. 12.62(a). If M_1 and M_2 operate as ideal current sources, then one PA's signal has little effect on the other's.[10] However, it is difficult to achieve a high efficiency while keeping M_1 and M_2 in saturation.

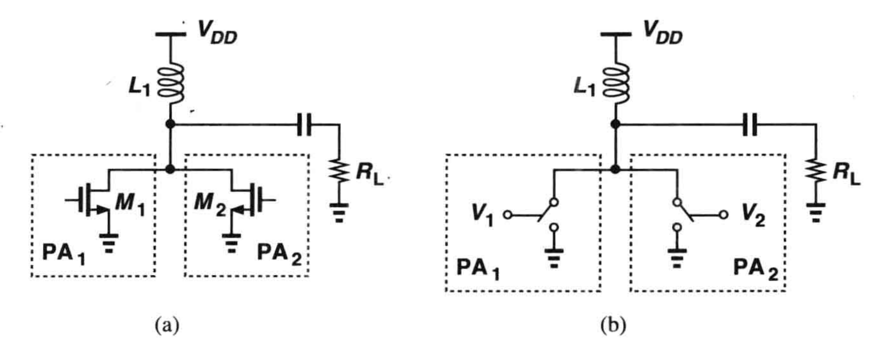


Figure 12.62 *(a) Example of combining circuit, (b) simple model.*

Now, suppose M_1 and M_2 enter the deep triode region and can be modeled as voltage-controlled switches [Fig. 12.62(b)]. In this case, the load seen by one PA is *modulated* by the other and hence varies with time, distorting the signal.

To formulate the interaction between the PAs, we consider the more common arrangement depicted in Fig. 12.63(a), where a transformer sums the outputs[11] and drives the load resistance. The output network can be simplified as shown in Fig. 12.63(b). We wish to determine the impedance seen by each PA with respect to ground. To this end, we must

10. The coupling of the output to the gate of each transistor through C_{GD} does create some interaction.

11. In this case, the transformer in fact *subtracts* V_2 from V_1. Thus, V_2 must be negated before reaching PA_2.

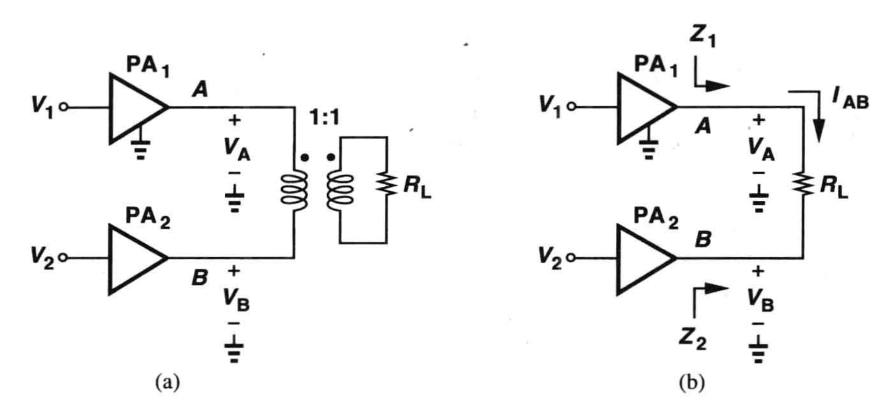


Figure 12.63 *(a) Outphasing with a transformer, (b) equivalent circuit.*

compute $I_{AB} = (V_A - V_B)/R_L$ and then $Z_1 = V_A/I_{AB}$ and $Z_2 = -V_B/I_{AB}$. If each PA stage is modeled as an ideal *voltage* buffer with a unity gain, then $V_A = V_1$ and $V_B = V_2$, yielding

$$I_{AB}(t) = \frac{V_1(t) - V_2(t)}{R_L} \tag{12.116}$$

$$= \frac{V_0 \sin(\omega_0 t + \phi + \theta) - V_0 \sin(\omega_0 t + \phi - \theta)}{2R_L} \tag{12.117}$$

$$= \frac{V_0 \cos(\omega_0 t + \phi) \sin\theta}{R_L}. \tag{12.118}$$

It follows that

$$\frac{V_A(t)}{I_{AB}(t)} = \frac{V_0 \sin(\omega_0 t + \phi)\cos\theta + V_0 \cos(\omega_0 t + \phi)\sin\theta}{2V_0 \cos(\omega_0 t + \phi)\sin\theta} R_L \tag{12.119}$$

$$= \frac{R_L}{2} + \frac{R_L}{2}\frac{\sin(\omega_0 t + \phi)}{\cos(\omega_0 t + \phi)} \cdot \theta. \tag{12.120}$$

We now assume θ is relatively *constant* with time, and transform this result to the frequency domain. Since the numerator and denominator of the fraction in the second term are 90° out of phase, they introduce a factor of $-j$ in the equivalent impedance. Thus,

$$Z_1 = \frac{R_L}{2} - j\cot\theta\frac{R_L}{2}; \tag{12.121}$$

i.e., the equivalent impedance seen by PA_1 consists of a real part equal to $R_L/2$ and an imaginary part equal to $(-\cot\theta)R_L/2$.[12] Similarly,

$$Z_2 = \frac{R_L}{2} + j\cot\theta\frac{R_L}{2}. \tag{12.122}$$

12. If the input waveforms are represented by cosines, the imaginary part is given by $(-\tan\theta)R_L/2$.

Example 12.29

It is often said that the reactive parts in Eqs. (12.121) and (12.122) correspond to capacitance and inductance, respectively. Is this statement accurate?

Solution:

Generally, it is not. Capacitive and inductive reactances must be proportional to frequency, whereas the second terms in Eqs. (12.121) and (12.122) are not. However, for a narrowband signal, a negative reactance can be viewed as a capacitance and a positive reactance as an inductance.

The dependence of Z_1 and Z_2 upon θ reveals that, if the PAs are *not* ideal voltage buffers, then the signal experiences a time-varying voltage division [Fig. 12.64(a)] and hence distortion. Recognized by Chireix [20], this effect can be alleviated if an additional reactance with opposite polarity is tied to each PA's output so as to cancel the second term in Eqs. (12.121) or (12.122) [Fig. 12.64(b)]. Since a parallel reactance (admittance) is usually preferred, we first transform Z_1 and Z_2 to admittances. Inverting the left-hand side of (12.121) and multiplying the numerator and denominator by $1 + j\cos\theta$, we have

$$Y_1 = \frac{2}{R_L}(\sin^2\theta + j\sin\theta\cos\theta). \tag{12.123}$$

To cancel the second term,

$$\frac{1}{j\omega_0 L_A} = -\frac{2}{R_L}j\sin\theta\cos\theta \tag{12.124}$$

and hence

$$L_A = \frac{R_L}{\omega_0\sin 2\theta}. \tag{12.125}$$

Similarly,

$$Y_2 = \frac{2}{R_L}(\sin^2\theta - j\sin\theta\cos\theta). \tag{12.126}$$

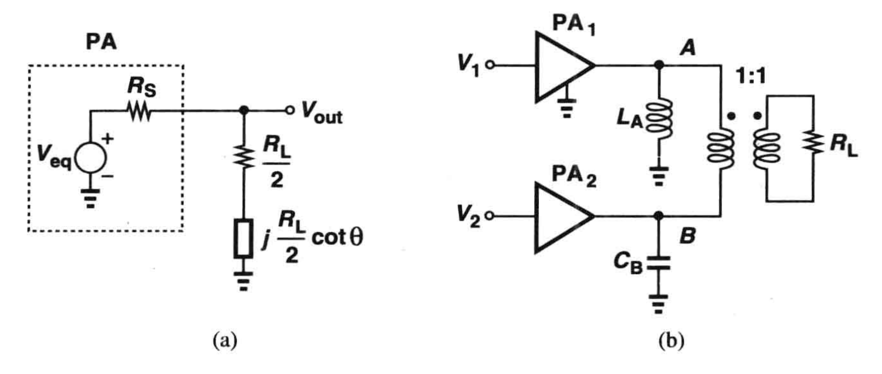


Figure 12.64 *(a) Time-varying voltage division in outphasing, (b) Chireix's cancellation technique.*

To cancel the second term in (12.122),

$$jC_B\omega_0 = \frac{2}{R_L}j\sin\theta\cos\theta, \tag{12.127}$$

and hence

$$C_B = \frac{\sin 2\theta}{R_L\omega_0}. \tag{12.128}$$

With perfect cancellation, $Z_1 = Z_2 = R_L/(2\sin^2\theta)$. Interestingly, L_A and C_B resonate at the carrier frequency because

$$L_AC_B = \frac{1}{\omega_0^2}. \tag{12.129}$$

The foregoing results are based on two assumptions: each PA can be approximated by a voltage source, and θ is relatively constant. The reader may view both suspiciously. After all, a heavily-switching PA stage exhibits an output impedance that swings between a small value (when the transistor is in the deep triode region) and a large value (when the transistor is off). Moreover, the envelope time variation translates to a time-varying θ. In other words, addition of a constant inductance and a constant capacitance to the output nodes provides only a rough compensation.

The reader may wonder if it is possible to construct a three-port power network that provides *isolation* between two of the ports, thereby avoiding the above interaction. It can be shown that such a network inevitably suffers from loss.

In order to improve the compensation, the inductance and capacitance can *track* the envelope variation [24]. However, since it is difficult to vary the inductance, we must seek an arrangement that lends itself to only capacitance variation. To this end, let us implement Chireix's cancellation technique as shown in Fig. 12.65(a). Interestingly, L_A and C_B shift

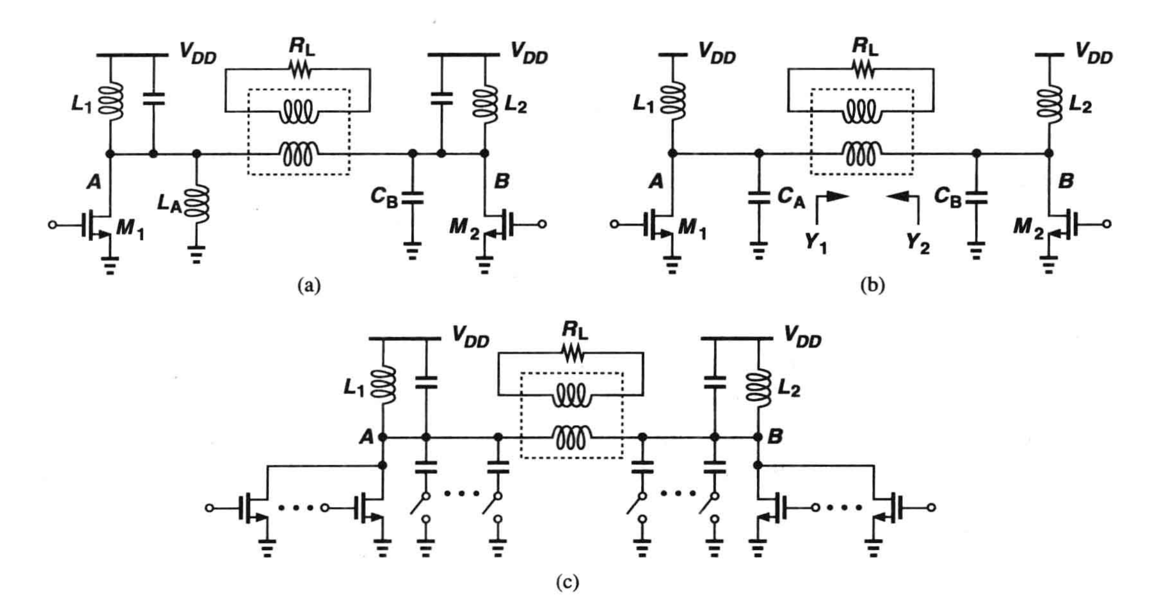


Figure 12.65 *(a) Outphasing PA using Chireix's technique, (b) addition of variable capacitances, (c) circuit with discrete capacitor arrays.*

the resonance frequencies of the two output tanks in *opposite* directions. We therefore surmise that if only unequal capacitors are tied to A and B and varied in opposite directions, then cancellation may still occur. As depicted in Fig. 12.65(b), we select C_A and C_B as [24]

$$C_A = C_0 + \Delta C \tag{12.130}$$

$$C_B = C_0 - \Delta C, \tag{12.131}$$

seeking the proper value of ΔC. The admittances of the tanks are given by

$$Y_{tank,A} = \frac{1}{jL_0\omega} + j(C_0 + \Delta C)\omega \tag{12.132}$$

$$Y_{tank,B} = \frac{1}{jL_0\omega} + j(C_0 - \Delta C)\omega, \tag{12.133}$$

where $L_1 = L_2 = L_0$. Noting that, for a narrowband signal, $1/(jL_0\omega)$ and $jC_0\omega$ cancel, we use Eqs. (12.123) and (12.132) to write the total admittance at A:

$$Y_{tot,A} = Y_{tank,A} + Y_1 \tag{12.134}$$

$$= j\Delta C\omega + \frac{2\sin^2\theta}{R_L} + \frac{j\sin 2\theta}{R_L}. \tag{12.135}$$

The reactive parts cancel if

$$\Delta C = -\frac{\sin 2\theta}{R_L\omega}. \tag{12.136}$$

Similarly, for node B:

$$Y_{tot,B} = Y_{tank,B} + Y_2 \tag{12.137}$$

$$= -j\Delta C\omega + \frac{2\sin^2\theta}{R_L} - j\frac{\sin 2\theta}{R_L}, \tag{12.138}$$

yielding the same ΔC as in (12.136), a fortunate coincidence.

The above development indicates that if ΔC varies in proportion to $\sin 2\theta$, then the cancellation is more accurate, leaving a real part in the overall impedance equal to

$$Re\{Y_{tot,B}\} = \frac{2\sin^2\theta}{R_L}. \tag{12.139}$$

Unfortunately, this component also varies with the envelope.[13] This issue can be alleviated by adjusting the strength of each PA so as to maintain a relatively constant output power [24]. Figure 12.65(c) shows the result [24], where both the capacitors and the transistors can be tuned in discrete steps. Utilizing bond wires for inductors and an off-chip balun, the PA delivers an output of 13 dBm in the WCDMA mode with a drain efficiency of 27% [24].

13. If the input waveforms are represented by cosines, then this real part is given by $2\cos^2\theta/R_L$.

12.9 DOHERTY POWER AMPLIFIER

The amplifier stages studied thus far incorporate a single output transistor, inevitably approaching saturation as the transistor enters the triode region (saturation region for bipolar devices). We therefore postulate that if an auxiliary transistor is introduced that provides gain only when the main transistor begins to compress, then the overall gain can remain relatively constant for higher input and output levels. Figure 12.66(a) illustrates this principle: the main amplifier remains linear for input swings up to about V_1, and the auxiliary amplifier contributes to the output power as the input exceeds V_1. The former operates in class A and the latter in class C.

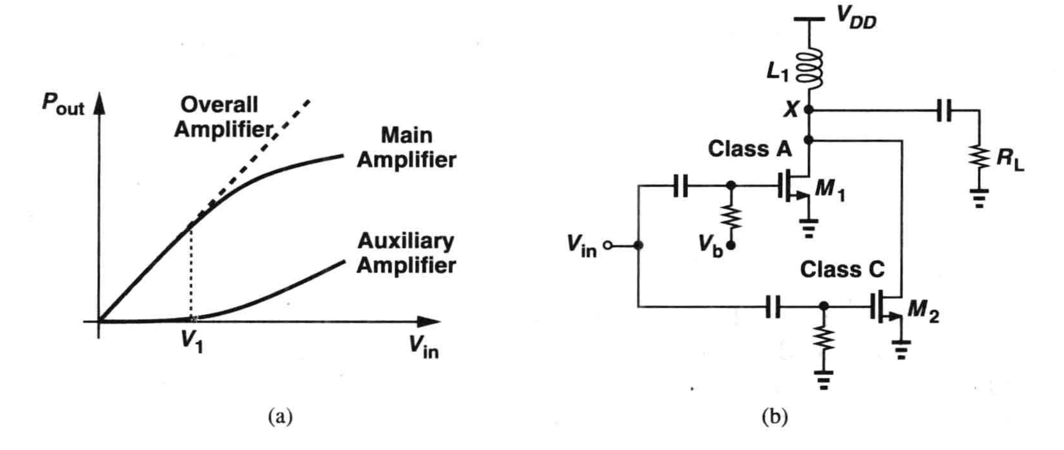


Figure 12.66 *(a) Input/output characteristics of a Doherty PA, (b) hypothetical implementation.*

While simple and elegant, the above principle is not straightforward to implement: How exactly should the auxiliary amplifier be tied to the main amplifier? Figure 12.66(b) shows an example where the currents produced by the two branches are simply summed at the output node. However, if the voltage swing at X is large enough to drive M_1 into the triode region, then it is likely to drive M_2 into the triode region, too.

Recognizing that amplitude-modulated signals reach their peak values only occasionally and hence cause a low average efficiency, Doherty has introduced the above two-path principle and developed the PA topology shown in Fig. 12.67(a) [25]. He has called the main and auxiliary stages the "carrier" and "peaking" amplifiers, respectively. The carrier PA is followed by a transmission line of length equal to $\lambda/4$, where λ denotes the carrier wavelength. To match the delay through this line, another $\lambda/4$ T-line is inserted in series with the input of the peaking amplifier.

In order to understand the operation of the Doherty PA, we construct the equivalent circuit shown in Fig. 12.67(b), where I_1 and I_2 represent the RF currents produced by the carrier and peaking stages, respectively. Our first objective is to determine the impedance Z_1. The voltage and current waveforms at a point x along a lossless transmission line are

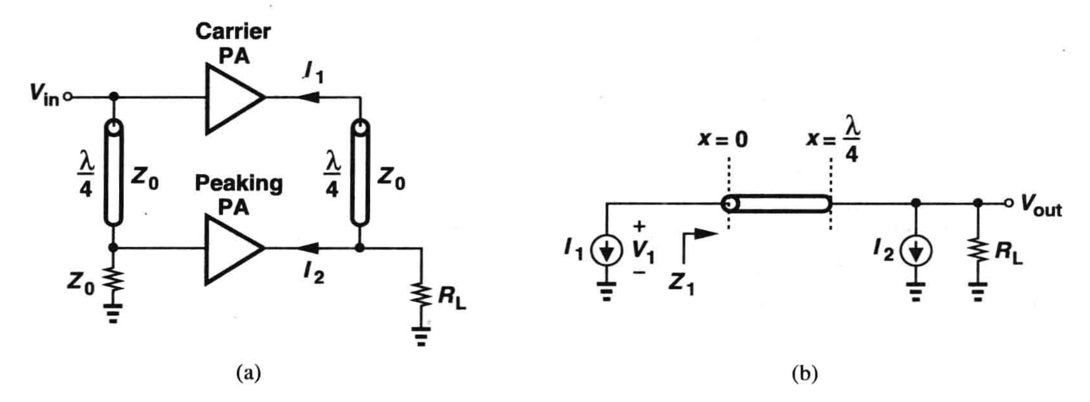


Figure 12.67 *(a) Conceptual realization of Doherty PA, (b) equivalent output network.*

respectively given by

$$V(t,x) = V^+ \cos(\omega_0 t - \beta x) + V^- \cos(\omega_0 t + \beta x) \tag{12.140}$$

$$I(t,x) = \frac{V^+}{Z_0}\cos(\omega_0 t - \beta x) - \frac{V^-}{Z_0}\cos(\omega_0 t + \beta x), \tag{12.141}$$

where the first term in each expression represents a wave propagating in the positive x direction and the second, a wave propagating in the negative x direction, $\beta = 2\pi/\lambda$, and Z_0 is the line's characteristic impedance. Since I_2 is delayed with respect to I_1 by $\lambda/4$ ($= 90^\circ$), we write $I_1 = I_0 \cos\omega_0 t$ and $I_2 = \alpha I_0 \cos(\omega_0 t - 90^\circ) = -\alpha I_0 \sin\omega_0 t$, where α is a proportionality factor signifying the relative "strength" of the peaking stage. Equations (12.140) and (12.141) must now be satisfied at $x = 0$:

$$V(t,0) = (V^+ + V^-)\cos\omega_0 t = V_1 \tag{12.142}$$

$$I(t,0) = \left(\frac{V^+}{Z_0} - \frac{V^-}{Z_0}\right)\cos\omega_0 t = -I_1, \tag{12.143}$$

and at $x = \lambda/4$:

$$V\left(t, \frac{\lambda}{4}\right) = (-V^+ + V^-)\sin\omega_0 t = V_{out} \tag{12.144}$$

$$I\left(t, \frac{\lambda}{4}\right) = \left(-\frac{V^+}{Z_0} - \frac{V^-}{Z_0}\right)\sin\omega_0 t. \tag{12.145}$$

Writing a KCL at the output node, we have

$$\frac{V_{out}}{R_L} + I_2 = I\left(t, \frac{\lambda}{4}\right), \tag{12.146}$$

and hence

$$\frac{(-V^+ + V^-)\sin\omega_0 t}{R_L} - \alpha I_0 \sin\omega_0 t = \left(-\frac{V^+}{Z_0} - \frac{V^-}{Z_0}\right)\sin\omega_0 t. \tag{12.147}$$

It follows that

$$\frac{V^+ - V^-}{R_L} + \alpha I_0 = \frac{V^+ + V^-}{Z_0}. \tag{12.148}$$

In the last step, we observe that $Z_1 = -V_1/I_1$, which from Eqs. (12.142) and (12.143) emerges as

$$Z_1 = -\frac{V^+ + V^-}{V^+ - V^-} Z_0. \tag{12.149}$$

Also, (12.143) yields $V^+ - V^- = -I_0 Z_0$ and hence $Z_1 = -(V^+ + V^-)/I_0$. Substituting these values in (12.148) gives

$$-\frac{I_0 Z_0}{R_L} + \alpha I_0 = -\frac{I_0 Z_1}{Z_0}, \tag{12.150}$$

and

$$Z_1 = Z_0 \left(\frac{Z_0}{R_L} - \alpha \right). \tag{12.151}$$

The key point here is that, as the peaking stage begins to amplify (α rises above zero), the load impedance seen by the main PA *falls*. This effect counteracts the increase of the main PA drain voltage swings that would be necessary for larger input levels, resulting in a relatively constant drain voltage swing beyond the transition point (Fig. 12.68). One can therefore choose V_1 such that the main PA operates in its linear region even for $V_{in} > V_1$.

Several properties of the Doherty PA can be derived [25]. We state the results here: (1) the technique extends the linear range by approximately 6 dB; (2) the efficiency reaches a theoretical maximum of 79% at full output power; (3) this efficiency is obtained if Z_0 in Fig. 12.67(a) is chosen equal to $2R_L$.

The Doherty PA presents its own challenges with respect to IC design. The two transmission lines, especially that at the output, introduce considerable loss, degrading the efficiency. Also, for large swings, the transistor in the peaking stage turns on and off, producing discontinuities in the derivatives of the output current and possibly yielding a high adjacent channel power. In other words, the circuit may prove useful if signal compression must be avoided but not if ACPR must remain small.

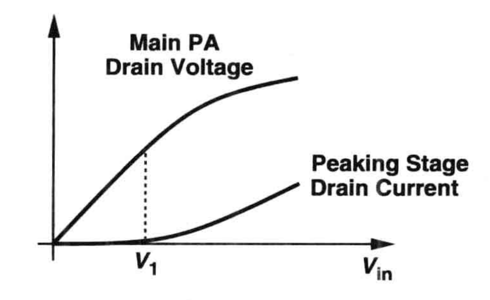


Figure 12.68 *Current and voltage variation in a Doherty PA.*

12.10 DESIGN EXAMPLES

Most power amplifiers employ two (or sometimes three) stages, with matching networks placed at the input, between the stages, and at the output (Fig. 12.69). The "driver" can be viewed as a buffer between the upconverter and the output stage, providing gain and driving the low input impedance of the latter. For example, if a PA must deliver +30 dBm, the two stages in Fig. 12.69 may have a gain of 25 to 30 dB, allowing the upconverter output to be in the range of 0 to + 5 dBm. Depending on the carrier frequency and the power levels, the first matching network, N_1, may be omitted, i.e., the driver simply senses the upconverter output voltage.

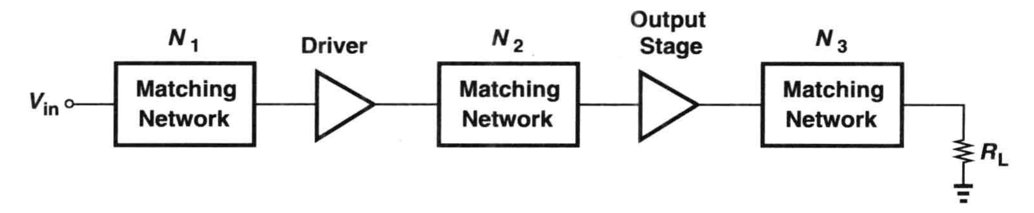


Figure 12.69 *Typical two-stage PA.*

The input and output matching networks in Fig. 12.69 serve different purposes: N_1 may provide a 50-Ω input impedance, whereas N_3 amplifies the voltage swings produced by the output stage (or, equivalently, transforms R_L to a lower value). The 50-Ω input impedance is necessary if the PA is designed as a stand-alone circuit that interfaces with the preceding circuit by means of external components. In an integrated TX, on the other hand, the upconverter/PA interface impedance can be chosen quite higher.

The matching network, N_2, in Fig. 12.69 is incorporated for practical reasons. Since the design may begin with load-pull measurements on the output transistor, the source impedance that this device must see for maximum efficiency is known and fixed once the design of the output stage is completed. Thus, the driver must drive such an input impedance, often requiring a matching network. In other words, the use of N_2 affords a modular design: first the output stage, next the driver, and last the interstage matching, with some iteration at the end. Without N_2, the driver and the output stage must be treated as a single circuit and co-designed for optimum performance. While possibly more complex, such a procedure may offer a somewhat higher efficiency because it avoids the loss of N_2.

In this section, we study a number of PA designs reported in the literature. As we will see, the efficiency and linearity vary substantially from one design to another. The reader is therefore cautioned that the comparison of the performance of different PAs is not straightforward. In particular, one must ask the following questions:

- What carrier frequency and maximum output power are targeted? The higher these are, the tighter the efficiency-linearity trade-off is.
- How much gain does the PA provide? Designs with lower gains tend to be more linear.
- Does the PA employ off-chip components? Most output matching networks are realized externally to avoid the loss of on-chip devices. For example, some designs

incorporate bond wires as part of this network—even though such PAs may be called "fully integrated."

- Does the IC technology provide thick metallization? For frequencies up to tens of gigahertz, a thick metal lowers the loss of on-chip inductors and transmission lines. (At higher frequencies, skin effect becomes dominant and the benefits of thick metalization diminish.)
- Does the design stress the transistor(s)? Many reported PAs employ a supply voltage equal to the maximum tolerable device voltage, V_{max}, but allow above-supply swings, possibly stressing the transistor(s).
- In what type of package is the PA tested? The package parasitics play a critical role in the performance of the PAs.
- Are the efficiency and ACPR measured at the same output power level? Some designs may quote the efficiency at the maximum power but the ACPR at a lower average output.

12.10.1 Cascode PA Examples

Nonlinear PAs can utilize cascode devices to reduce the stress on transistors. Figure 12.70 shows a class E example for the 900-MHz band [26]. Here, M_3 and M_4 turn on for part of the input swing. The use of a cascode device affords nearly twice the drain voltage swing (compared to a simple common-source stage), allowing the load resistance at the drain to be quadrupled. Consequently, the matching network need only transform 50 Ω to about 4.4 Ω for an output power of 1 W, exhibiting smaller losses. For these power levels, the on-resistance of the M_1–M_2 branch is chosen to be about 1.2 Ω, smaller than other equivalent resistances in the matching network, but requiring a W/L of 15 mm/0.25 μm for each! The large drain capacitance of M_2 is absorbed in C_1, and the gate capacitance of M_1 is tuned by a 2-nH bond wire and an external variable capacitance. Inductors L_2 and L_3 are also realized by bond wires.

The input stage consisting of M_3 and M_4 in Fig. 12.70 operates as a class C amplifier because the transistors have a negligible bias current until the swing raises V_B above

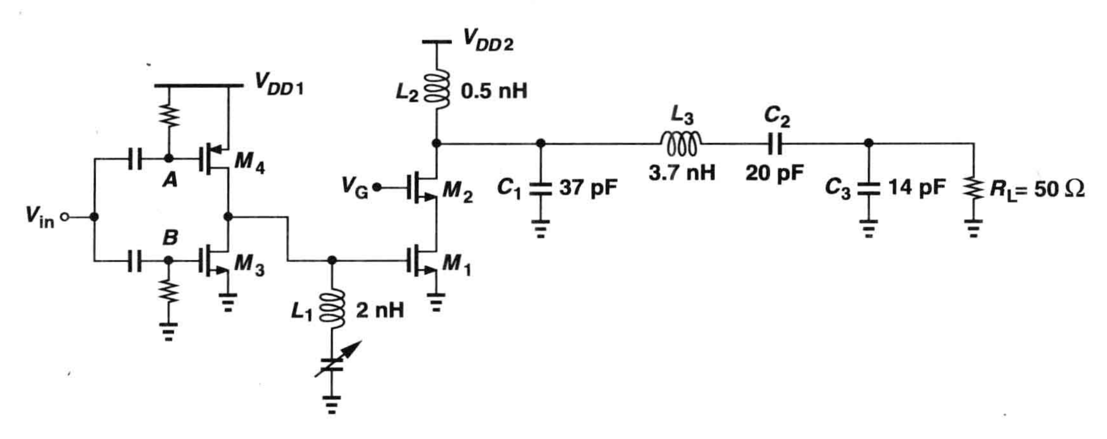


Figure 12.70 *Class E PA example.*

V_{TH3} or drops V_A below $V_{DD} - |V_{TH4}|$. The PA achieves a power-added efficiency of 41% while delivering 0.9 W with $V_{DD1} = 2.5$ V and $V_{DD2} = 1.8$ V. The actual design employs two copies of the circuit in quasi-differential form and combines the outputs by means of an off-chip balun [26].

Figure 12.71(a) shows another example of cascode PAs [27]. In order to allow even larger swings at the drain of M_2, this topology bootstraps the gate of the cascode device to the output through R_1. In other words, since V_P and hence V_Q rise with V_{out}, M_2 now experiences less stress than if V_P were constant. Of course, if V_P tracks V_{out} with unity gain, then M_2 operates as a diode-connected device, limiting the minimum value of V_{out}.[14] For this reason, capacitor C_1 is added, creating a fraction of the output swing at V_P. Figure 12.71(b) plots the circuit's waveforms, revealing that the maximum drain-source voltages experienced by M_1 and M_2 can be made approximately equal [27], leading to a large tolerable output swing.

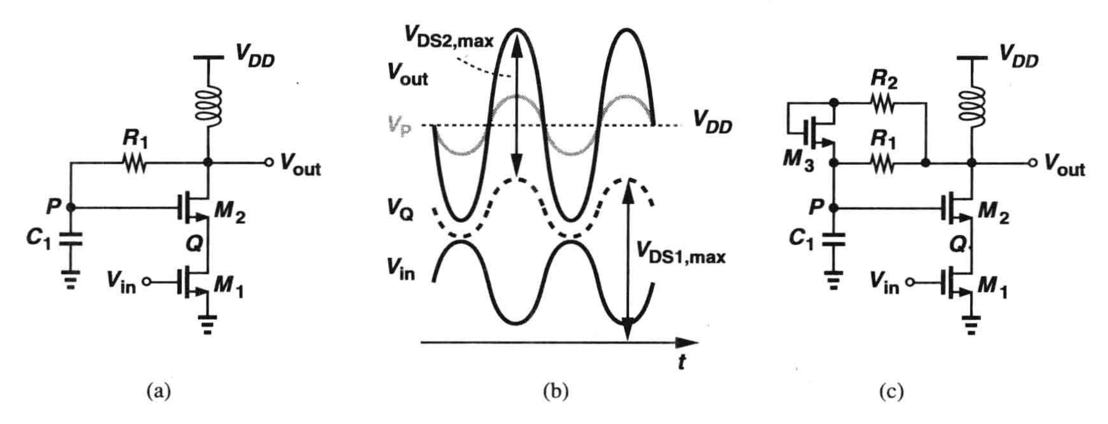


Figure 12.71 *(a) Cascode PA with bootstrapping, (b) circuit's waveforms, (c) addition of diode-connected device.*

Example 12.30

In the ideal case, what output voltage swing does the topology of Fig. 12.71(a) provide?

Solution:

In the ideal case, V_{DD} can be chosen equal to the maximum allowable drain-source voltage, V_{max}, so that V_{out} can swing from nearly zero to about $2V_{DD} = 2V_{max}$. This is possible if at $V_{out} = 2V_{max}$, the gate voltage of M_2 is raised enough to yield $V_{DS2} = V_{DS1} = V_{max}$.

The topology of Fig. 12.71(a) can be further improved by making the bootstrap path somewhat unilateral so that the positive swings are *larger* than the negative swings. Depicted in Fig. 12.71(c), the modified circuit includes an additional series branch consisting of R_2 and a diode-connected device, M_3. As V_{out} rises, M_3 turns on, allowing the

14. And forfeiting the benefits of cascode operation.

gate voltage of M_2 to follow. On the other hand, as V_{out} falls, M_3 turns off, and only R_1 can pull the gate down.

Example 12.31

Explain what happens to the output duty cycle in the presence of asymmetric positive and negative swings.

Solution:

Since the swing above V_{DD} is larger than that below, the duty cycle must be less than 50% to yield an average voltage still equal to V_{DD}. The average output *power* nonetheless increases. This can be seen from the nearly ideal waveforms shown in Fig. 12.72, where we have

$$V_1 T_1 \approx V_2 T_2 \tag{12.152}$$

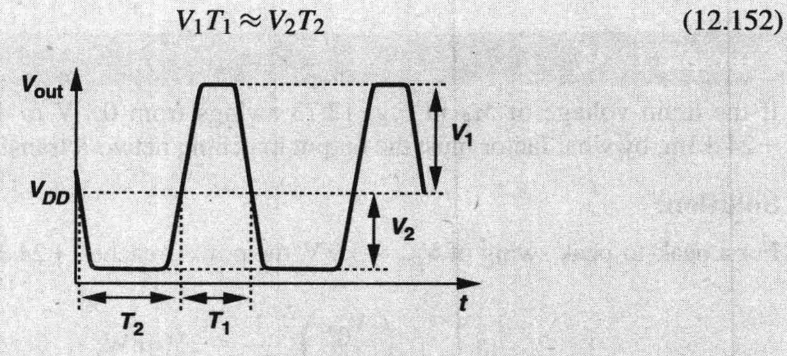


Figure 12.72 *Bootstrapped cascode waveforms in the presence of asymmetric swings.*

to ensure the average voltage is equal to V_{DD}. The average power is given by

$$P_{avg} \approx \frac{(V_1 + V_2)^2 T_1}{T_1 + T_2}, \tag{12.153}$$

which, from Eq. (12.152), reduces to

$$P_{avg} \approx \left(1 + \frac{T_2}{T_1}\right) V_2^2. \tag{12.154}$$

Thus, as V_1 increases and hence T_1 decreases, P_{avg} rises because $V_2 \approx V_{DD}$.

Figure 12.73 shows the overall bootstrapped cascode PA design for the 2.4-GHz band [27]. The dashed box encloses the on-chip circuitry, L_1–L_3 denote bond wires, and T_1–T_7 are transmission lines implemented as traces on the printed-circuit board. The output stage utilizes device widths of $W_3 = 2$ mm and $W_4 = 1.5$ mm (with $L = 0.18\ \mu$m), presenting an input capacitance of roughly 4 pF. In the driver stage, $W_1 = 600\ \mu$m and $W_2 = 300\ \mu$m.

The circuit employs three matching networks: (1) T_1, C_1, and T_2 match the input to 50 Ω; (2) T_3, L_2, and C_2 provide interstage matching; and (3) L_3, T_4–T_6, C_3, and C_4 transform the 50-Ω load to a lower resistance. Transmission line T_7 acts as an open circuit at 2.4 GHz.

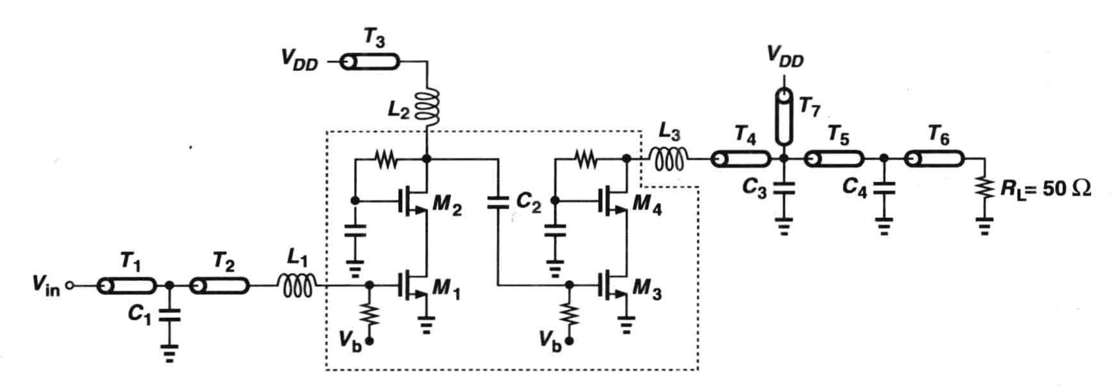


Figure 12.73 *Implementation of bootstrapped PA.*

Example 12.32

If the drain voltage of M_4 in Fig. 12.73 swings from 0.1 V to 4 V and the PA delivers +24 dBm, by what factor must the output matching network transform the load resistance?

Solution:

For a peak-to-peak swing of $V_{pp} = 3.9$ V, the power reaches +24 dBm (=250 mW) if

$$\left(\frac{V_{pp}}{2\sqrt{2}}\right)^2 \frac{1}{R_{in}} = 250\,\text{mW}, \tag{12.155}$$

where R_{in} is the resistance seen at the drain of M_4. It follows that

$$R_{in} = 7.6\ \Omega. \tag{12.156}$$

The output matching network must therefore transform the load by a factor of 6.6.

Operating with a supply of 2.4 V, the PA of Fig. 12.73 delivers a maximum (saturated) output of 24.5 dBm with a gain of 31 dB and a PAE of 49%. The output 1-dB compression is around 21 dBm.

Another example of cascode PA design is conceptually illustrated in Fig. 12.74(a) [28]. Here, a class B stage is added in parallel with a class A amplifier, contributing gain as the latter begins to compress. The operation is similar to that shown in Fig. 12.66(a) for the Doherty PA. The summation of the two outputs faces the same issue illustrated in Fig. 12.66(b), but if the two stages experience compression at the *input*, then their outputs can be simply summed in the current domain [28]. From this assumption emerges the PA circuit shown in Fig. 12.74(b), where M_1–M_4 form the main class A stage and M_5–M_6 the class B path. In this design, $(W/L)_{1,2} = 192/0.8$, $(W/L)_{3,4} = 1200/0.34$, and $(W/L)_{5,6} = 768/0.18$ (all dimensions are in microns). Note that $(W/L)_{5,6} > (W/L)_{1,2}$ because the class B devices take over at high output levels. The cascode transistors have a thicker oxide and longer channel so as to allow a higher voltage swing at the output.

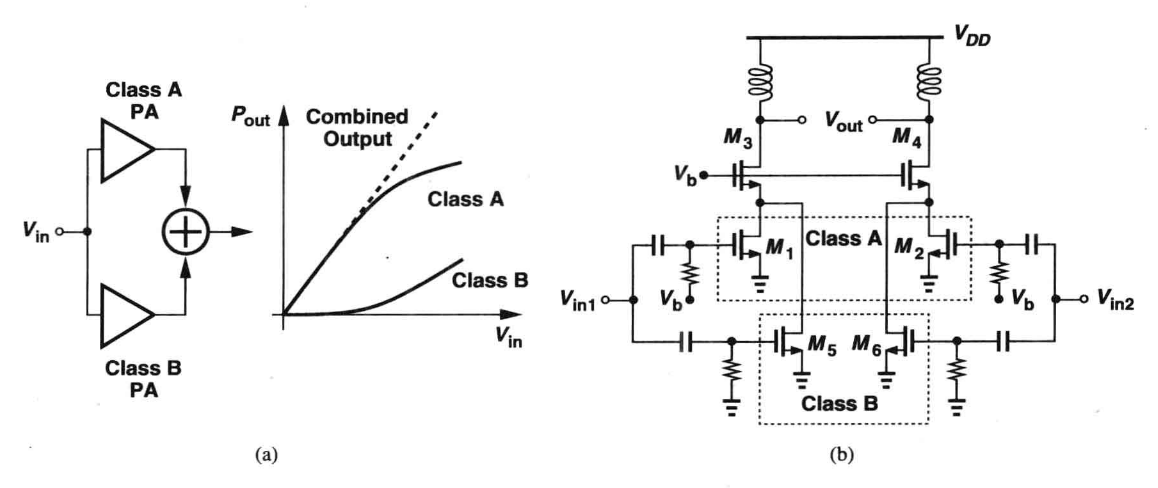


Figure 12.74 *(a) Parallel class A and B PAs to raise compression point, (b) realization of circuit.*

The PA of Fig. 12.74(b) produces a maximum output of 22 dBm with a PAE of 44%. The small-signal gain is 12 dB and the output P_{1dB} is 20.5 dBm.[15]

12.10.2 Positive-Feedback PAs

Our study of PAs in this chapter has revealed relatively large output transistors and the difficulty in driving them by the preceding stage. Now suppose, as conceptually illustrated in Fig. 12.75(a), the output transistor is decomposed into two, and one device, M_2, is driven by an inverted copy of V_{out} rather than by V_{in}. The input capacitance of the stage is therefore reduced proportionally. The implementation of the idea becomes straightforward in a differential design [Fig. 12.75(b)]. Since the input devices can now be substantially smaller, they are more easily switched, leading to a higher efficiency.

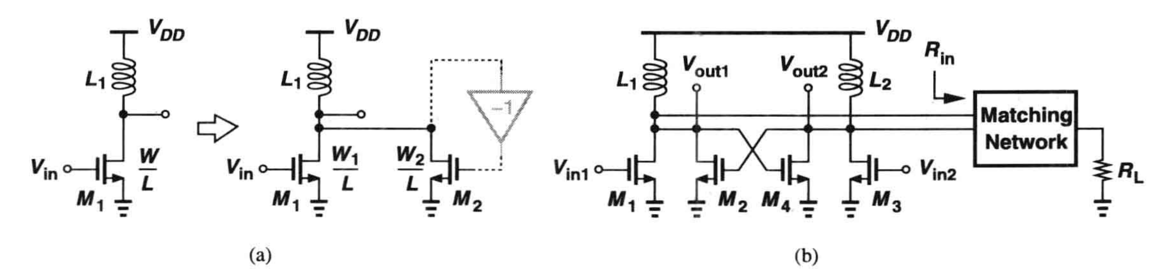


Figure 12.75 *(a) Decomposition of an output device with one section driven by the output, (b) PA driving its own capacitance.*

How should the drive capability be partitioned between M_1–M_3 and M_2–M_4 in Fig. 12.75(b)? We are tempted to allocate most of the required width to M_2–M_4 so as to minimize W_1 and W_3. However, as the design is skewed in this direction, two effects

15. The operation frequency and the supply voltage are not mentioned. It is unclear which components are external.

manifest themselves: (1) The capacitance at the output node becomes so large that it may dictate a small resonating inductance (L_1 and L_2) and hence a low output power. This issue is less problematic in class E stages where the output capacitance can be absorbed in the matching network. (2) As M_2 and M_4 become wider and carry a proportionally higher current, they form an oscillator with L_1 and L_2, which are loaded by the equivalent resistance, R_{in}.

Is it possible to employ an oscillatory PA stage? For a variable-envelope signal, such a circuit would create considerable distortion. However, for a constant-envelope waveform, an oscillatory stage may prove acceptable if its output phase can faithfully track the input phase. In other words, the cross-coupled oscillator must be injection-locked to the input with sufficient bandwidth so that the input phase excursions travel to the output unattenuated. If M_1 and M_3 in Fig. 12.75(b) are excessively small with respect to M_2 and M_4, then the input coupling factor may not guarantee locking. Of course, the lock range must be wide enough to cover the entire transmit band. In particular, the lock range can be expressed as

$$\Delta\omega = \pm\frac{\omega_0}{2Q}\frac{g_{m1,3}}{g_{m2,4}}, \tag{12.157}$$

where $Q \approx L_{1,2}\omega/(R_{in}/2)$. With a typical R_{in} of a few ohms, the lock range is usually quite wide.

Figure 12.76 shows a 1.9-GHz class E PA based on injection locking [29]. Both stages incorporate positive feedback, and the inductors are realized by bond wires. In this design, all transistors have a channel length of 0.35 μm, W_5–$W_8 = 980\,\mu$m, $W_1 = W_3 = 3600\,\mu$m, and $W_2 = W_4 = 4800\,\mu$m. Also, L_1–$L_4 = 0.37$ nH, $L_5 = L_6 = 0.8$ nH, and $C_D = 5.1$ pF. A microstrip balun on the PCB converts the differential output to single-ended form.

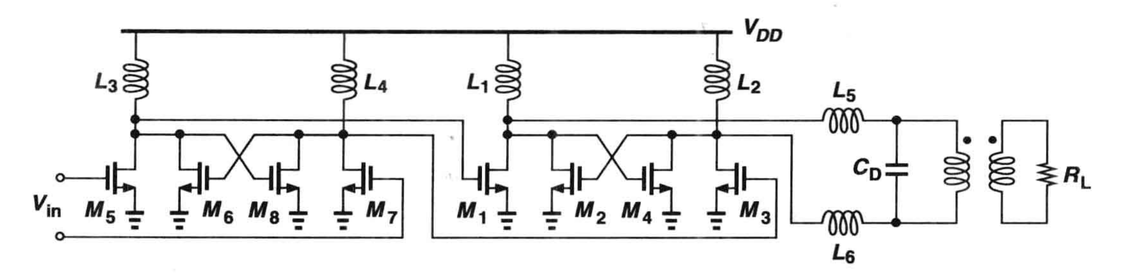


Figure 12.76 *Injection-locked PA example.*

Operating with a 2-V supply and producing a maximum drain voltage of 5 V, the circuit of Fig. 12.76 delivers 1 W of power with a PAE of 48%. It is suited to constant-envelope modulation schemes such as GMSK.

An interesting issue here relates to output power control. While in other topologies, reduction of the input level eventually produces an arbitrarily small output (even if the circuit is nonlinear), injection-locked PAs deliver a relatively large output even if the input amplitude falls to zero (if the circuit oscillates). Figure 12.77 depicts an example where M_p controls the bias current of the output stage. However, to ensure negligible efficiency degradation at the maximum output level, the on-resistance of this device with $V_{cont} \approx 0$ must be very small, requiring a very wide transistor.

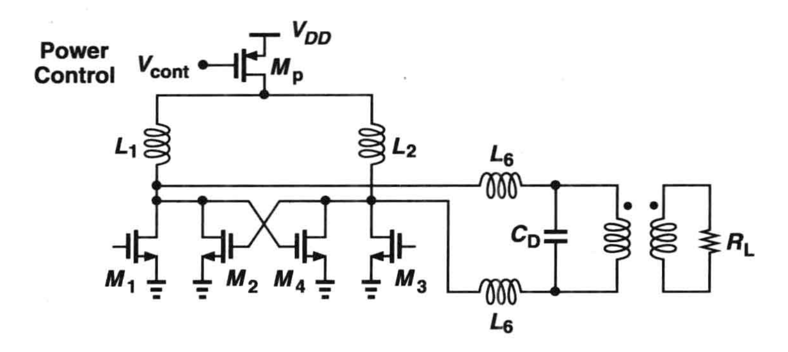


Figure 12.77 *Injection-locked PA with output power control.*

12.10.3 PAs with Power Combining

We have observed in this chapter that transistor stress issues limit the supply voltage and hence output swing of PAs, dictating a matching network with a large impedance transformation ratio. We may alternatively ask, is it possible to directly add the output *voltages* of several stages so as to generate a large output power.

Let us return to the notion of transformer-based matching [Fig. 12.78(a)]. The on-chip realization of 1-to-n transformers poses many difficulties, especially if the primary and/or secondary must carry large currents. For example, both the series resistance and the inductance of the primary must be kept very small if power levels of greater than hundreds of milliwatts are to be delivered. Also, as explained in Chapter 7, stacked transformers contain various parasitics, and multi-turn planar transformers can hardly achieve a turns ratio of greater than 2. In other words, it is desirable to employ only 1-to-1 transformers.

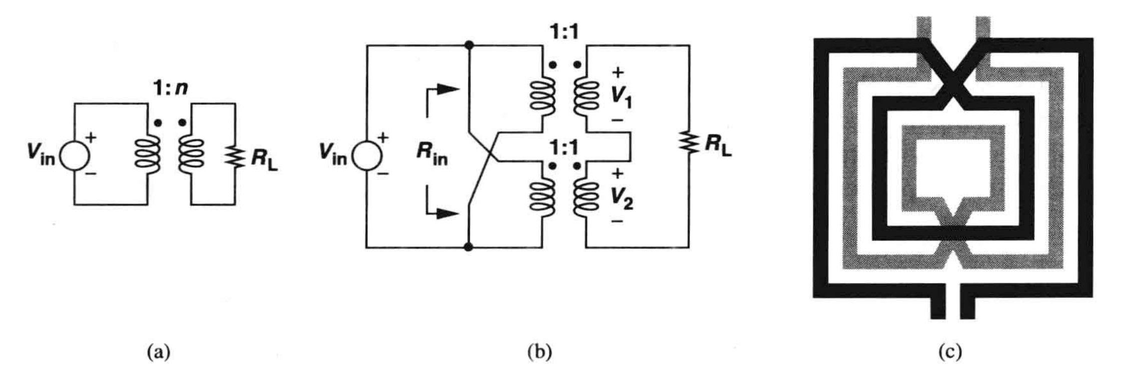


Figure 12.78 *(a) Output stage model using a 1-to-n transformer, (b) circuit using two 1-to-1 transformers to combine the outputs, (c) simple 1-to-1 transformer.*

With these issues in mind, we pursue transformer-based matching but using the approach shown in Fig. 12.78(b). Here, the primaries of two 1-to-1 transformers are placed in *parallel* while their secondaries are tied in *series* [30]. We expect that the circuit amplifies the voltage swing by a factor of 2 because $V_1 = V_2 = V_{in}$. As exemplified by Fig. 12.78(c), 1-to-1 transformers more easily lend themselves to integration.

Example 12.33

Determine the equivalent resistance seen by V_{in} in Fig. 12.78(b) if the transformer loss is neglected.

Solution:

Since the power delivered to R_L is $P_{out} = (2V_{in})^2/R_L$, where V_{in} denotes the rms value of the input, we have

$$P_{in} = P_{out} \tag{12.158}$$

$$= \frac{4V_{in}^2}{R_L}. \tag{12.159}$$

Also, $P_{in} = V_{in}^2/R_{in}$, yielding

$$R_{in} = \frac{R_L}{4}, \tag{12.160}$$

which is identical to that of a 1-to-2 transformer driving a load resistance of R_L.

How is an actual output stage connected to the double-transformer topology of Fig. 12.78(b)? We can envision the simple arrangement depicted in Fig. 12.79(a), but the long, high-current-carrying interconnects between the amplifier and the two primaries introduce loss and additional inductance. Alternatively, we can "slice" the amplifier into two equal sections and place each in the close vicinity of its respective primary [Fig. 12.79(b)]. In this case, the amplifier *input* lines may be long, a less serious issue because they carry smaller currents.

The concept illustrated in Fig. 12.79(b) can be extended to a multitude of 1-to-1 transformers so as to obtain a greater R_L/R_{in} ratio. Figure 12.80 shows a 2.4-GHz class E example employing four differential branches [30]. Each inductor is realized as an on-chip straight, wide metal line to handle large currents with a small resistance. For class E operation, a capacitor must be placed between the drains of each two input (differential)

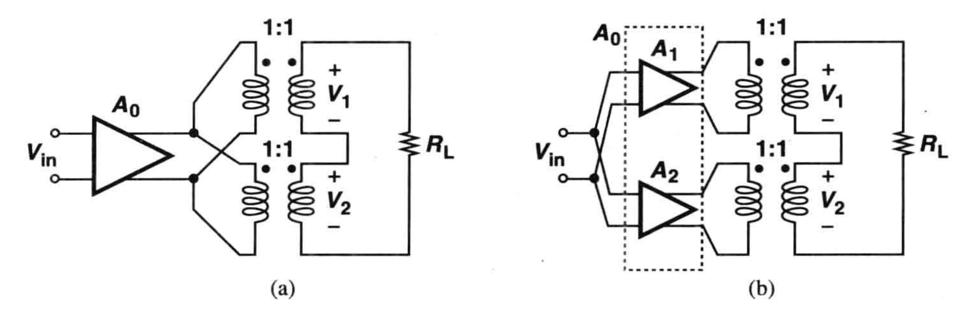


Figure 12.79 *(a) A single PA or (b) two PAs driving two transformers.*

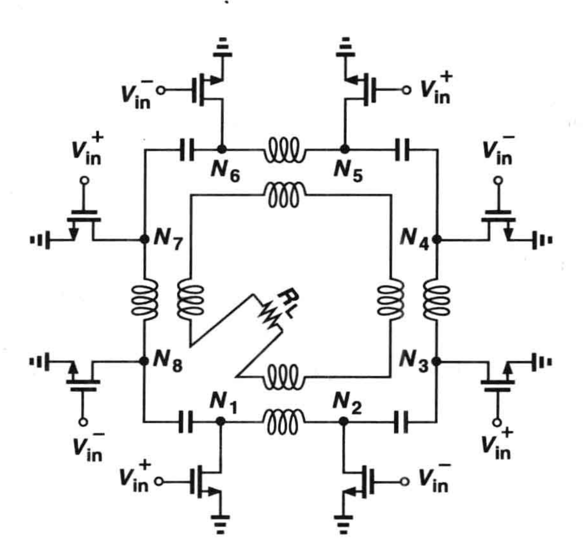


Figure 12.80 *Power combining technique in [30].*

transistors, but the physical distance between N_1 and N_2, etc., inevitably adds inductance in series with the capacitor. Since the odd-numbered nodes in Fig. 12.80 have the same potential, and so do the even-numbered nodes, the capacitor is tied between, for example, N_2 and N_3 rather than between N_1 and N_2.

Example 12.34

Determine the differential resistance seen by each amplifier in Fig. 12.80 if the transformers are lossless.

Solution:

Returning to the simpler case illustrated in Fig. 12.79(b), we recognize that each of A_1 and A_2 sees twice the resistance seen by A_0, i.e., $R_L/2$. Thus, for the four-amplifier arrangement of Fig. 12.80, each differential pair sees a load resistance of $R_L/4$.

Designed for a 2-W output level [30], the circuit of Fig. 12.80 incorporates wide input transistors. To create input matching, inductors are inserted between V_{in}^- and V_{in}^+ of adjacent branches. The differential inputs are first routed to the center of the secondary and then distributed to all four amplifiers, thus minimizing phase and amplitude mismatches. One factor limiting the efficiency of transformer-based PAs is the primary/secondary coupling factor, typically no higher than 0.6 for planar structures [30].

The design in Fig. 12.80 is realized in 0.35-μm technology with a 3-μm thick top metal layer, producing an output of 1.9 W (32.8 dBm) with a PAE of 41%. The PA provides a small-signal gain of 16 dB and runs from a 2-V supply. The output P_{1dB} is around 27 dBm.

Example 12.35

The gain of the above PA falls to 8.7 dB at full output power [30]. Estimate the power consumed by a stage necessary to drive this PA.

Solution:

The driver must deliver 32.8 dBm−8.7 dB = 24.1 dBm (= 257 mW). From previous examples, such a power can be obtained with an efficiency of about 40%, translating to a power consumption of about 640 mW. Since the above PA draws approximately 4 W from the supply,[16] we note that the driver would require an additional 16% power consumption.

The multiple amplifiers driving the 1-to-1 transformers in the foregoing topologies can also be turned off individually, thus allowing output power control [31]. As illustrated in Fig. 12.81, if only M of the N amplifiers are on, then the output voltage swing drops by a factor of N/M. The notable benefit of this approach is that, as the output power is scaled down, it provides a higher efficiency than conventional PAs [31]. [The primary of the off stage(s) must be shorted by a switch.]

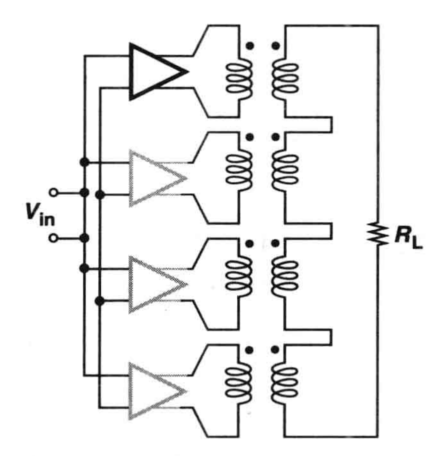


Figure 12.81 *Power combining with switchable stages.*

It is also possible to place the secondaries of the transformers in parallel so as to add their output currents [32].

12.10.4 Polar Modulation PAs

As explained in Section 12.7, a critical issue in polar modulation is the design of the supply modulation circuit for minimum degradation of efficiency and headroom. Figure 12.82 shows an example of an envelope path [33]. Here, a "delta modulator" (DM) generates a replica of V_{env} at the V_{DD} node of the PA output stage. The DM loop consists of a comparator, a buffer, and a low-pass filter.[17] Owing to the high gain of the comparator, the loop

16. The drain efficiency is 48% [30].

17. A zero must be added to this loop to ensure stability [33].

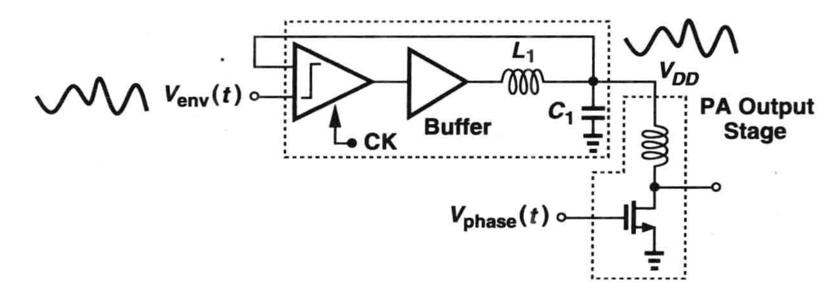


Figure 12.82 *Polar modulation PA using a delta modulator for envelope path.*

ensures that the average output tracks the input even though the comparator produces only a binary waveform.

In the circuit of Fig. 12.82, the output stage's average current flows through the LPF and the buffer. To minimize loss of efficiency and headroom, the LPF utilizes an (off-chip) inductor rather than a resistor, and the buffer must employ very wide transistors. Moreover, the DM loop bandwidth must accommodate the envelope signal spectrum and introduce a delay that can be matched by the phase path.

Figure 12.83 shows an example of a polar modulation transmitter [19]. In contrast to the topologies studied in Section 12.7, this architecture merges the envelope and phase loops: the highly-linear cascade of MX_1 and VGA_1 downconverts and reproduces both components at an IF, and the decomposition occurs at this IF. The output power is controlled by means of VGA_1 and VGA_2, e.g., as their gain increases, so does the output level such that the envelope at B remains equal to that at A. This also guarantees that the swing delivered to the feedback limiter is constant and it can be optimized for minimum AM/PM conversion. This transmitter consists of several modules realized in BiCMOS and GaAs technologies. The system delivers an output of +29 dBm in the EDGE mode at 900 MHz [19].

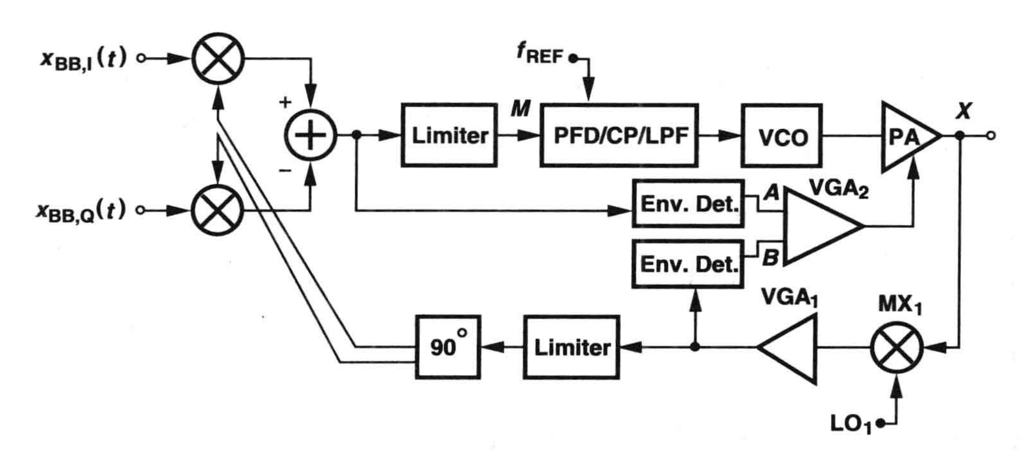


Figure 12.83 *Polar modulation PA with envelope and phase feedback.*

Depicted in Fig. 12.84(a) is another polar transmitter [18]. Here, the quadrature upconverter operates independently, generating an IF waveform having both envelope and phase components. The two signals are then extracted, with the former controlling the output stage and the latter driving an offset PLL.

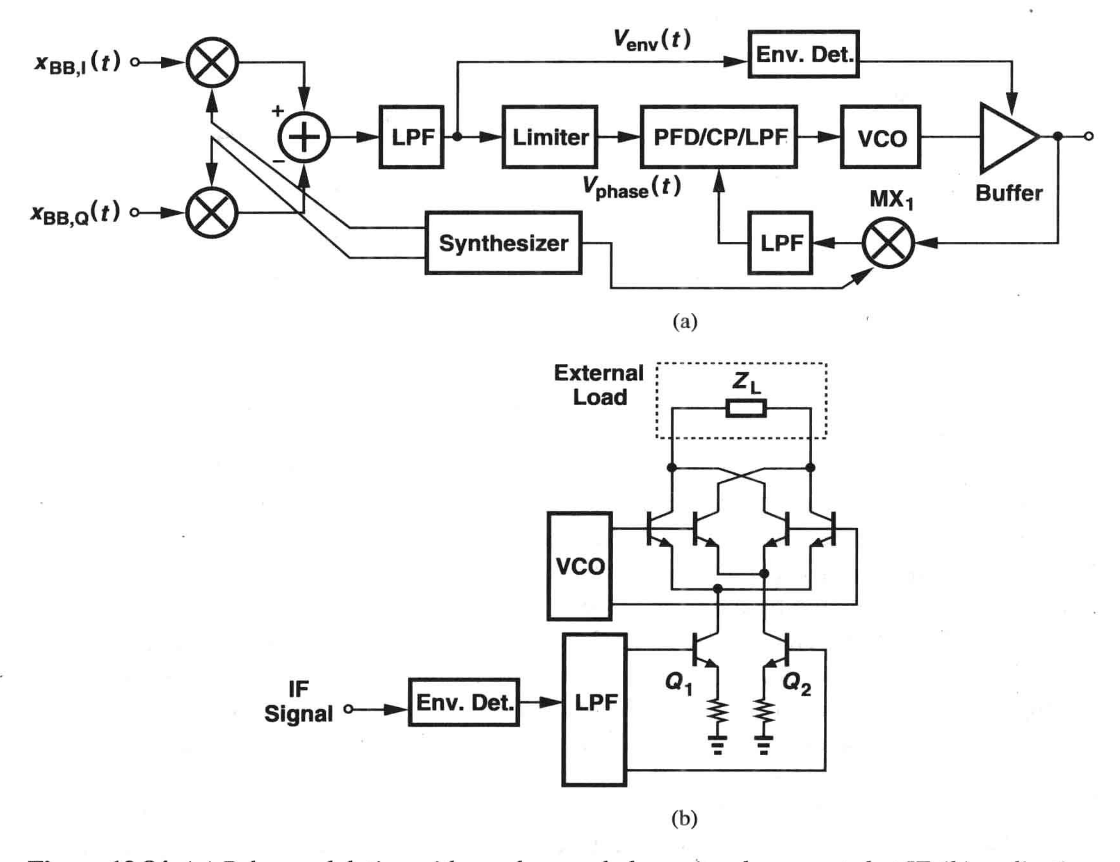


Figure 12.84 *(a) Polar modulation with envelope and phase signals separated at IF, (b) realization of output combining circuit.*

Figure 12.84(b) shows the details of the TX front end. It consists of an envelope detector, a low-pass filter, and a double-balanced mixer driven by the VCO. Designed to deliver a power of +1 dBm, the mixer multiplies the envelope by the phase signal produced by the VCO, thus generating the composite waveform at the output [18]. As mentioned in Section 12.7, the dc offset in the envelope path leads to leakage of the phase component; this TX employs offset cancellation in the envelope path to suppress this effect.

The reader may wonder why the polar transmitters studied above do not employ a mixer of this type to combine the envelope and phase signals. Figure 12.84(b) suggests that the mixer requires a large voltage headroom, consuming substantial power. This technique is thus suited to low or moderate output levels.

12.10.5 Outphasing PA Example

Recall that outphasing transmitters incorporate two identical nonlinear PAs and sum their outputs to obtain the composite signal. Figure 12.85 shows the circuit realization of one PA for the 5.8-GHz band [34, 35]. An on-chip transformer serves as an input balun, applying differential phases to the driver stage. Inductors L_1 and L_2 and capacitors C_1 and C_2 provide interstage matching. The output stage operates in the class E mode, with

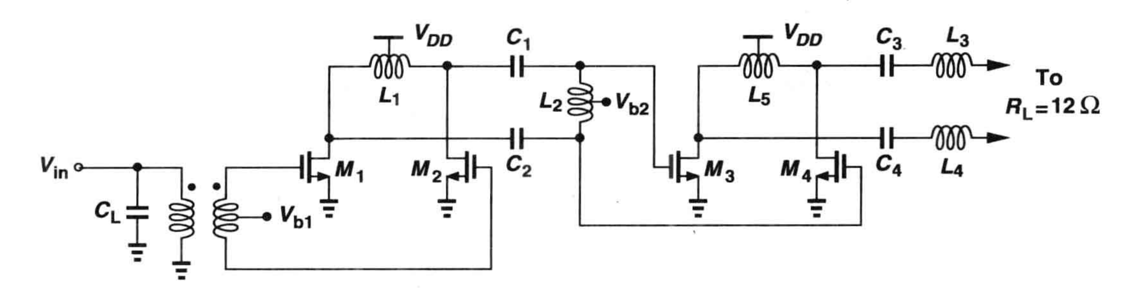


Figure 12.85 *PA used in an outphasing system.*

L_3–L_5 and C_3 and C_4 shaping the nonoverlapping voltage and current waveforms. Note that the design assumes a load resistance of 12 Ω, a value provided by the power combiner described below.

Example 12.36

If the above circuit operates with a 1.2-V supply and the minimum drain voltage is 0.15 V, estimate the peak drain voltage of M_3 and M_4.

Solution:

We note from Section 12.3.2 that the peak drain voltage is roughly equal to $3.56V_{DD} - 2.56V_{DS}$. Thus, the drain voltage reaches 3.9 V. In the actual design, the peak drain voltage is 3.5 V [34, 35].

Example 12.37

If the circuit of Fig. 12.85 delivers a power of 15.5 dBm to the 12-Ω load [34, 35], compare the drain voltage swing with that across R_L.

Solution:

Since 15.5 dBm corresponds to 35.5 mW, the peak-to-peak differential voltage swing across R_L is equal to $2\sqrt{2}\sqrt{(35.5\text{ mW})R_L} = 1.85$ V. Thus, the class-E output network in fact *reduces* the voltage swing by a factor of 3.8 in this case.[18] From a device stress point of view, this is undesirable.

In order to sum the outputs of the PAs, the outphasing TX employs a "Wilkinson combiner" rather than a transformer. Recall from Section 12.3.2 that a transformer ideally exhibits no loss but it allows interaction between the two PAs. By contrast, a Wilkinson combiner ideally provides isolation between the two input ports but suffers from loss.

18. Of course, the drain signal contains stronger harmonics than the output signal does.

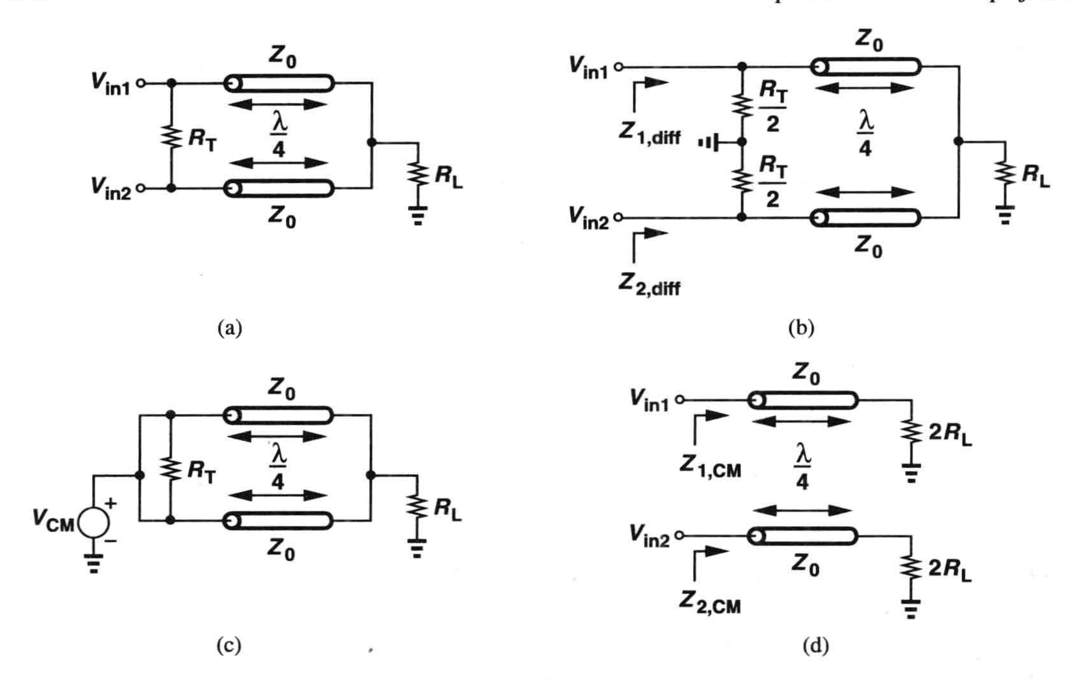


Figure 12.86 *(a) Wilkinson power combiner, (b) equivalent circuit with differential inputs, (c) equivalent circuit with a common-mode input, (d) input CM impedance.*

Shown in Fig. 12.86(a), the combiner consists of two quarter-wavelength transmission lines and a resistor, R_T.

The Wilkinson divider is commonly analyzed in terms of "odd" (differential) and "even" (common-mode) inputs. For differential inputs in Fig. 12.86(a), the output summing junction and the midpoint of R_T are at ac ground [Fig. 12.86(b)]. The $\lambda/4$ lines transform the short circuit to an open circuit, yielding

$$Z_{1,diff} = Z_{2,diff} = \frac{R_T}{2}. \tag{12.161}$$

That is, the differential component of V_{in1} and V_{in2} causes dissipation in R_T but not in R_L. For a common-mode input, all the points in the circuit rise and fall in unison [Fig. 12.86(c)]. Thus, R_L can be replaced with two parallel resistors of value $2R_L$, and R_T with an open circuit [Fig. 12.86(d)]. In this case, the impedance seen by each voltage source is given by

$$Z_{1,CM} = Z_{2,CM} = \frac{Z_0^2}{2R_L}. \tag{12.162}$$

We recognize that the common-mode component of V_{in1} and V_{in2} causes dissipation in R_L but not in R_T.

Example 12.38

How does the Wilkinson combiner of Fig. 12.86(a) achieve isolation between the input ports?

Solution:

If the impedance seen by each input voltage source is constant and independent of differential or common-mode components, then V_{in1} does not "feel" the presence of V_{in2} and vice versa. This condition is satisfied if

$$Z_{1,diff} = Z_{1,CM} \tag{12.163}$$

$$Z_{2,diff} = Z_{2,CM}. \tag{12.164}$$

Denoting all of these impedances by Z_{in}, we write

$$Z_{in} = \frac{R_T}{2} = \frac{Z_0^2}{2R_L}. \tag{12.165}$$

The result expressed by Eq. (12.162) reveals that the Wilkinson combiner can also transform the load impedance to a desired value if Z_0 is chosen properly. The outphasing system in [34, 35] transforms $R_L = 50\,\Omega$ to $Z_{in} = 12\,\Omega$ using $Z_0 = 35\,\Omega$. The combining of the two differential PA outputs requires four transmission lines, each having a length of 2.8 mm. The on-chip lines are wrapped around the PA circuitry and realized as shown in Fig. 12.87.

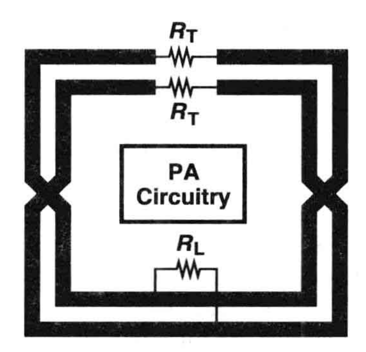


Figure 12.87 *On-chip Wilkinson combiner used at the output of outphasing system.*

Designed in 0.18-μm technology, the outphasing PA of Fig. 12.85 incorporates thick-oxide transistors to sustain a peak drain voltage of 3.5 V. The overall circuit generates an output of 18.5 dBm with an efficiency of 47% while amplifying a 64-QAM OFDM signal.

REFERENCES

[1] S. Cripps, *RF Power Amplifiers for Wireless Communications,* Norwood, MA: Artech House, 1999.

[2] A. Grebebbikov, *RF and Microwave Power Amplifier Design,* Boston: McGraw-Hill, 2005.

[3] A. Johnson, "Physical Limitations on Frequency and Power Parameters of Transistors," *RCA Review,* vol. 26, pp. 163–177, 1965.

[4] A. A. Saleh, "Frequency-Independent and Frequency-Dependent Nonlinear Models of TWT Amplifiers," *IEEE Tran. Comm.,* vol. COM-29, pp. 1715–1720, Nov. 1981.

[5] C. Rapp, "Effects of HPA-Nonlinearity on a 4-DPSK/OFDM-Signal for a Digital Sound Broadband System," *Rec. Conf. ECSC,* pp. 179–184, Oct. 1991.

[6] J. C. Pedro and S. A. Maas, "A Comparative Overview of Microwave and Wireless Power-Amplifier Behavioral Modeling Approaches," *IEEE Tran. MTT,* vol. 53, pp. 1150–1163, April 2005.

[7] H. L. Kraus, C. W. Bostian, and F. H. Raab, *Solid State Radio Engineering,* New York: Wiley, 1980.

[8] S. C. Cripps, "High-Efficiency Power Amplifier Design," presented in Short Course: RF ICs for Wireless Communication, Portland, June 1996.

[9] J. Staudinger, "Multiharmonic Load Termination Effects on GaAs MESFET Power Amplifiers," *Microwave J.* pp. 60–77, April 1996.

[10] N. O. Sokal and A. D. Sokal, "Class E - A New Class of High-Efficiency Tuned Single-Ended Switching Power Amplifiers," *IEEE J. of Solid-State Circuits,* vol. 10, pp. 168–176, June 1975.

[11] F. H. Raab, "An Introduction to Class F Power Amplifiers," *RF Design*, pp. 79–84, May 1996.

[12] H. Seidel, "A Microwave Feedforward Experiment," *Bell System Technical J.*, vol. 50, pp. 2879–2916, Nov. 1971.

[13] E. E. Eid, F. M. Ghannouchi, and F. Beauregard, "Optimal Feedforward Linearization System Design," *Microwave J.*, pp. 78–86, Nov. 1995.

[14] D. P. Myer, "A Multicarrier Feedforward Amplifier Design," *Microwave J.*, pp. 78–88, Oct. 1994.

[15] R. E. Myer, "Nested Feedforward Distortion Reduction System," US Patent 6127889, Oct., 2000.

[16] L. R. Kahn, "Single-Sideband Transmission by Envelope Elimination and Restoration," *Proc. IRE*, vol. 40, pp. 803–806, July 1952.

[17] W. B. Sander, S. V. Schell, and B. L. Sander, " Polar Modulator for Multi-Mode Cell Phones," *Proc. CICC*, pp. 439–445, Sept. 2003.

[18] M. R. Elliott et al., "A polar modulator transmitter for GSM/EDGE," *IEEE J. of Solid-State Circuits,* vol. 39, pp. 2190–2199, Dec. 2004.

[19] T. Sowlati et al., "Quad-band GSM/GPRS/EDGE Polar Loop Transmitter," *IEEE J. of Solid-State Circuits,* vol. 39, pp. 2179–2189, Dec. 2004.

[20] H. Chireix, "High-Power Outphasing Modulation," *Proc. IRE,* pp. 1370–1392, Nov. 1935.

[21] D. C. Cox, "Linear Amplification with Nonlinear Components," *IEEE Tran. Comm.*, vol. 22, pp. 1942–1945, Dec. 1974.

[22] D. C. Cox and R. P. Leek, "Component Signal Separation and Recombination for Linear Amplification with Nonlinear Components," *IEEE Tran. Comm.*, vol. 23, pp. 1281–1287, Nov. 1975.

[23] F. J. Casadevall, "The LINC Transmitter," *RF Design*, pp. 41–48, Feb. 1990.

[24] S. Moloudi et al., "An Outphasing Power Amplifier for a Software-Defined Radio Transmitter," *ISSCC Dig. Tech. Papers,* pp. 568–569, Feb. 2008.

[25] W. H. Doherty, "A New High Efficiency Power Amplifier for Modulated Waves," *Proc. IRE,* vol. 24, pp. 1163–1182, Sept. 1936.

[26] C. Yoo and Q. Huang, "A Common-Gate Switched, 0.9 W Class-E Power Amplifier with 41% PAE in 0.25-μm CMOS," *VLSI Circuits Symp. Dig. Tech. Papers,* pp. 56–57, June 2000.

[27] T. Sowlati and D. Leenaerts, "2.4 GHz 0.18-μm CMOS Self-Biased Cascode Power Amplifier with 23-dBm Output Power," *IEEE J. of Solid-State Circuits,* vol. 38, pp. 1318–1324, Aug. 2003.

[28] Y. Ding and R. Harjani, "A CMOS High-Efficiency +22-dBm Linear Power Amplifier," *Proc. CICC,* pp. 557–560, Sept. 2004.

[29] K. Tsai and P. R. Gray, "A 1.9-GHz 1-W CMOS Class E Power Amplifier for Wireless Communications," *IEEE J. Solid-State Circuits,* vol. 34, pp. 962–970, 1999.

[30] I. Aoki et al., "Fully-Integrated CMOS Power Amplifier Design Using the Distributed Active Transformer Architecture," *IEEE J. Solid-State Circuits,* vol. 37, pp. 371–383, March 2002.

[31] G. Liu et al., "Fully Integrated CMOS Power Amplifier with Efficiency Enhancement at Power Back-Off," *IEEE J. Solid-State Circuits,* vol. 43, pp. 600–610, March 2008.

[32] A. Afsahi and L. E. Larson, "An Integrated 33.5 dBm Linear 2.4 GHz Power Amplifier in 65 nm CMOS for WLAN Applications," *Proc. CICC,* pp. 611–614, Sept. 2010.

[33] D. K. Su and W. J. McFarland, "An IC for Linearizing RF Power Amplifiers Using Envelope Elimination and Restoration," *IEEE J. Solid-State Circuits,* vol. 33, pp. 2252–2259, Dec. 1998.

[34] A. Pham and C. G. Sodini, "A 5.8-GHz 47% Efficiency Linear Outphase Power Amplifier with Fully Integrated Power Combiner," *IEEE RFIC Symp. Dig. Tech. Papers,* pp. 160–163, June 2006.

[35] A. Pham, *Outphasing Power Amplifiers in OFDM Systems,* PhD Dissertation, MIT, Cambridge, MA, 2005.

PROBLEMS

12.1. Following the derivations leading to Eq. (12.16), prove that the other 50% of the supply power is dissipated by the transistor itself.

12.2. In Fig. 12.16, plot the current from V_{DD} as a function of time. Does this circuit provide the benefits of differential operation? For example, is the bond wire inductance in series with V_{DD} critical?

12.3. Prove that in Fig. 12.17, the voltage swings above and below V_{DD} are respectively equal to $2\sqrt{2}I_pR_L/\pi$ and $\sqrt{2}(\pi-2)I_pR_L/\pi$, where I_p denotes the peak drain current. (Hint: the average value of V_X and V_Y must be equal to V_{DD}.)

12.4. From Example 12.11, sketch the scaling factor for the output transistor width as α varies from near zero to $\pi/2$.

12.5. Compute the maximum efficiency of the cascode PA shown in Fig. 12.31(a). Assume M_1 and M_2 nearly turn off but their drain currents can be approximated by sinusoids.

12.6. Assuming a third-order nonlinearity for the envelope detector in Fig. 12.46, prove that the output spectrum of the system exhibits growth in the adjacent channels.

12.7. Repeat the calculations leading to Eq. (12.77) but assuming that the phase signal experiences a delay mismatch of ΔT.

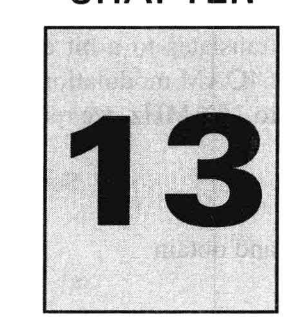
12.8. If transistor M_2 in Fig. 12.49(b) has an average current of I_0 and an average drain-source voltage of V_0, determine the efficiency of the stage. Neglect the on-resistance of M_1.

12.9. Derive Eq. (12.115) if $\theta(t) = \sin^{-1}[V_{env}(t)/V_1]$.

12.10. Does the Doherty amplifier of Fig. 12.67(a) operate properly if the input is driven by an ideal voltage source? Explain your reasoning.

12.11. In the Doherty amplifier of Fig. 12.67(a), the value of α is chosen equal to 0.5. Plot the waveforms at $x = 0$ and $x = \lambda/4$, assuming $Z_0 = R_L$.

CHAPTER



TRANSCEIVER DESIGN EXAMPLE

Having studied the principles of RF architecture and circuit design in the previous chapters, we are now prepared to embark upon the design of a complete transceiver. In this chapter, we design in 65-nm CMOS technology a dual-band transceiver for IEEE 802.11a/g applications. We first translate the standard's specifications to circuit design parameters and subsequently decide on the architecture and the frequency planning necessary to accommodate the 2.4-GHz and 5-GHz bands. The chapter outline is shown below.

System-Level Specifications	RX Design	TX Design	Synthesizer Design
▪ RX NF, IP_3, AGC, and I/Q Mismatch	▪ Broadband LNA	▪ PA	▪ VCO
▪ TX Output Power and P_{1dB}	▪ Passive Mixer	▪ Upconverter	▪ Dividers
▪ Synthesizer Phase Noise and Spurs	▪ AGC		▪ Charge Pump
▪ Frequency Planning			

In the circuit designs described in this chapter, the channel length of the transistors is equal to 60 nm unless otherwise stated. The reader is encouraged to review the 11a/g specifications described in Chapter 3.

13.1 SYSTEM-LEVEL CONSIDERATIONS

In deriving the transceiver specifications, we must bear two points in mind. (1) Since each nonideality degrades the performance to some extent, the budget allocated for each must leave sufficient margin for others. For example, if the RX noise figure is chosen to yield exactly the required sensitivity, then the I/Q mismatch may further degrade the BER. Thus, the overall performance must eventually be evaluated with *all* of the nonidealities present. (2) Both the TX and the RX corrupt the signal, dictating that each be designed with sufficient margin for the other's imperfections.

13.1.1 Receiver

For the receiver design, we must determine the required noise figure, linearity, and automatic gain control (AGC) range. In addition, we must decide on the maximum I and Q mismatch that can be tolerated in the downconverted signal.

Noise Figure As mentioned in Chapter 3, 11a/g specifies a packet error rate of 10%. This translates to a bit error rate of 10^{-5}, which in turn necessitates an SNR of 18.3 dB for 64QAM modulation [1]. Since TX baseband pulse shaping reduces the channel bandwidth to 16.6 MHz, we return to

$$\text{Sensitivity} = -174\,\text{dBm/Hz} + \text{NF} + 10\log\text{BW} + \text{SNR} \tag{13.1}$$

and obtain

$$\text{NF} = 18.4\,\text{dB} \tag{13.2}$$

for a sensitivity of −65 dBm (at 52 Mb/s). In practice, signal detection in the digital baseband processor suffers from nonidealities, incurring a "loss" of a few decibels. Moreover, the front-end antenna switch exhibits a loss of around 1 dB. For these reasons, and to deliver competitive products, manufacturers typically target an RX noise figure of about 10 dB. Since the 11a/g sensitivities are chosen to require about the same NF for different data rates, the NF of 10 dB must be satisfied for the highest sensitivity (−82 dBm) as well.

Nonlinearity For RX nonlinearity, we begin with the 1-dB compression point. As computed in Chapter 3, for 52 subchannels, the peak-to-average ratio reaches 9 dB,[1] requiring a P_{1dB} of at least −21 dBm so as to handle a maximum input level of −30 dBm. Allowing 2 dB for envelope variation due to baseband pulse shaping, we select a P_{1dB} of −19 dBm for the receiver. This value corresponds to an IIP_3 of about −9 dBm. However, the IIP_3 may also be dictated by adjacent channel specifications.

Let us examine the adjacent and alternate channel levels described in Chapter 3. At a sensitivity of −82 dBm, these levels are respectively 16 dB and 32 dB higher, and their intermodulation must negligibly corrupt the desired channel. We represent the desired, adjacent, and alternate channels by $A_0 \cos\omega_0 t$, $A_1 \cos\omega_1 t$, and $A_2 \cos\omega_2 t$, respectively. For a third-order nonlinearity of the form $y(t) = \alpha_1 x(t) + \alpha_2 x^2(t) + \alpha_3 x^3(t)$, the desired output is given by $\alpha_1 A_0 \cos\omega_0 t$ and the IM_3 component at ω_0 by $3\alpha_3 A_1^2 A_2/4$ (Fig. 13.1). The modulation scheme used with this sensitivity (BPSK) requires an SNR of 4 to 5 dB. Thus, we choose the IM_3 corruption to be around −15 dB to allow for other nonidealities:

$$20\log\left|\frac{3\alpha_3 A_1^2 A_2}{4\alpha_1 A_0}\right| = -15\,\text{dB}. \tag{13.3}$$

At this point, we can compute A_j as *voltage* quantities, substitute their values in the above equation, and determine $IIP_3 = \sqrt{|4\alpha_1|/|3\alpha_3|}$. Alternatively, we can maintain the

1. As explained in Section 13.3, this 9-dB "back-off" is quite conservative, leaving several decibels of margin for the TX nonlinearity.

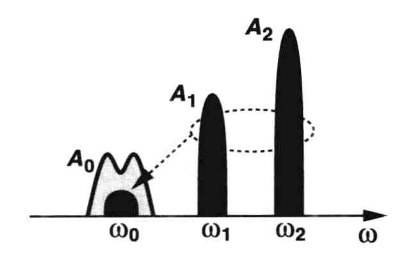


Figure 13.1 *Effect of intermodulation between two blockers.*

logarithmic quantities and proceed *very* carefully:

$$20\log\left|\frac{3\alpha_3}{4\alpha_1}\right| = -15\,\text{dB} - 40\log A_1 - 20\log A_2 + 20\log A_0. \tag{13.4}$$

Even though the last three terms on the right-hand side are voltage quantities, we replace them with their respective power levels in dBm:[2] $20\log A_1 = -63\,\text{dBm}$, $20\log A_2 = -47\,\text{dBm}$, and $20\log A_0 = -79$ dBm. It follows that

$$20\log\left|\frac{3\alpha_3}{4\alpha_1}\right| = +79\,\text{dBm}. \tag{13.5}$$

That is,

$$IIP_3|_{dBm} = 20\log\sqrt{\left|\frac{4\alpha_1}{3\alpha_3}\right|} \tag{13.6}$$

$$= -39.5\,\text{dBm}. \tag{13.7}$$

In Problem 13.1, we repeat this calculation for the data rate of 54 Mb/s and sensitivity of −65 dBm, obtaining roughly the same IIP_3. Thus, the IIP_3 value dictated by adjacent channel specifications is relatively relaxed in 11a/g. Of course, the baseband filters must still sufficiently attenuate the adjacent and alternate channels.

It is important to recognize the different design requirements related to the two IP_3 values obtained above. The IP_3 corresponding to the 1-dB compression point (sometimes called the "in-channel" IP_3) is satisfied if compression by the *desired* signal is avoided. This can be accomplished by lowering the receiver gain for high input levels. On the other hand, the IP_3 arising from adjacent channel specifications (sometimes called the "out-of-channel" IP_3) must be satisfied while the desired signal is only 3 dB above the reference sensitivity. In this case, the RX gain cannot be reduced to improve the linearity because the sensitivity degrades.

We now turn our attention to the IP_2 of the receiver. In this case, we are concerned with the demodulation of an interferer's envelope as a result of even-order nonlinearity. Since 64QAM OFDM interferers exhibit about 9 dB of peak-to-average ratio and hence a relatively "deep" amplitude modulation, this effect may appear particularly severe. However, as described in [2], the required IIP_2 is around 0 dBm, a value readily obtained in typical designs.

2. Recall from Chapter 3 that the desired input is at 3 dB above the reference sensitivity in this test.

AGC Range The receiver must automatically control its gain if the received signal level varies considerably. In order to determine the RX gain range, we consider both the 11a/g rate-dependent sensitivities described in Chapter 3 and the compression specification. The input level may vary from −82 dBm (for 6 Mb/s) to −65 dBm (for 54 Mb/s), and in each case the signal is amplified to reach the baseband ADC full scale, e.g., 1 V_{pp} (equivalent to +4 dBm in a 50-Ω system).[3] It follows that the RX gain must vary to accommodate the rate-dependent sensitivities. The challenge is to realize this gain range while maintaining a noise figure of about 10 dB (even at the lowest gain, for 54 Mb/s) and an (out-of-channel) IIP_3 of about −40 dBm (even at the highest gain, for 6 Mb/s).

Example 13.1

Determine the AGC range of an 11a/g receiver so as to accommodate the rate-dependent sensitivities.

Solution:

At first glance, we may say that the input signal level varies from −82 dBm to −65 dBm, requiring a gain of 86 dB to 69 dB so as to reach 1 V_{pp} at the ADC input. However, a 64QAM signal exhibits a peak-to-average ratio of about 9 dB; also, baseband pulse shaping to meet the TX mask also creates 1 to 2 dB of additional envelope variation. Thus, an average input level of −65 dBm in fact may occasionally approach a peak of −65 dBm + 11 dB = −54 dBm. It is desirable that the ADC digitize this peak without clipping. That is, for a −65-dBm 64QAM input, the RX gain must be around 58 dB. The −82-dBm BPSK signal, on the other hand, displays only 1 to 2 dB of the envelope variation, demanding an RX gain of about 84 dB.

The receiver gain range is also determined by the maximum allowable desired input level (−30 dBm). As explained in the above example, the baseband ADC preferably avoids clipping the peaks of the waveforms. Thus, the RX gain in this case is around 32 dB for BPSK (to raise the level from −30 dBm + 2 dB to +4 dBm) signals and 23 dB for 64QAM inputs (to raise the level from −30 dBm + 11 dB to +4 dBm). In other words, the RX gain must vary from (a) 84 dB to 58 dB with no NF degradation and an IIP_3 of −42 dBm (above example), and (b) from 58 dB to 23 dB with at most a dB-per-dB rise in the NF *and* at least a dB-per-dB rise in P_{1dB}.[4]

Figure 13.2 sketches the required RX behavior in terms of its gain, NF, and IIP_3 variation with the input signal level. The actual number of steps chosen here depends on the design of the RX building blocks and may need to be quite larger than that depicted in Fig. 13.2.

3. A 1-V_{pp} differential swing translates to a peak single-ended swing of 0.25 V, a reasonable value for a 1.2-V supply.

4. That is, for every dB of gain reduction, the NF must rise by no more than 1 dB.

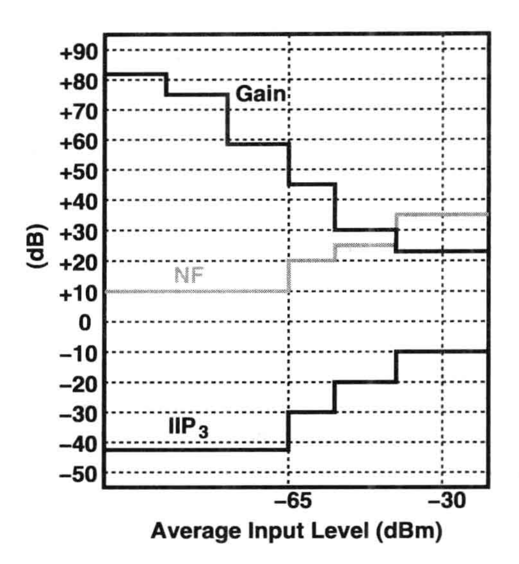


Figure 13.2 *Required RX gain switching and NF and IIP_3 variations.*

Example 13.2

The choice of the gain in the above example guarantees that the signal level reaches the ADC full scale for 64QAM as well as BPSK modulation. Is that necessary?

Solution:

No, it is not. The ADC resolution is selected according to the SNR required for 64QAM modulation (and some other factors studied in Section 13.2.3). For example, a 10-bit ADC exhibits an SNR of about 62 dB, but a BPSK signal can tolerate a much lower SNR and hence need not reach the ADC full scale. In other words, if the BPSK input is amplified by, say, 60 dB rather than 84 dB, then it is digitized with 6 bits of resolution and hence with ample SNR ($\approx$ 38 dB) (Fig. 13.3). In other words, the above AGC calculations are quite conservative.

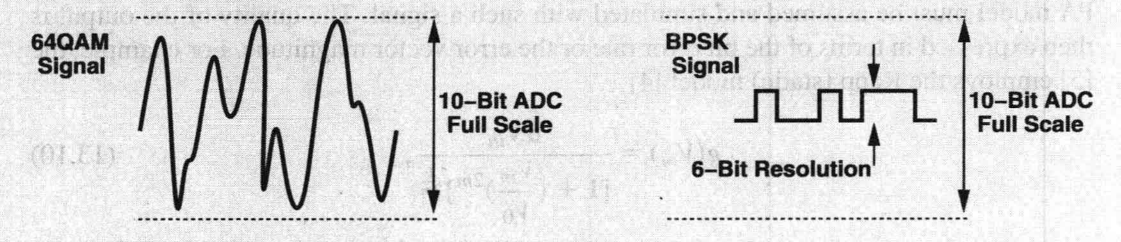


Figure 13.3 *Available ADC resolution for a full-scale signal and a smaller input swing.*

I/Q Mismatch The I/Q mismatch study proceeds as follows. (1) To determine the tolerable mismatch, we must apply in system simulations a 64QAM OFDM signal to a direct-conversion receiver and measure the BER or the EVM. Such simulations are

repeated for various combinations of amplitude and phase mismatches, yielding the acceptable performance envelope. (2) Using circuit simulations and random device mismatch data, we must compute the expected I/Q mismatches in the quadrature LO path and the downconversion mixers. (3) Based on the results of the first two steps, we must decide whether the "raw" matching is adequate or calibration is necessary. For 11a/g, the first step suggests that an amplitude mismatch of 0.2 dB and a phase mismatch of 1.5° are necessary [3]. Unfortunately such tight matching requirements are difficult to achieve without calibration.

Example 13.3

A hypothetical image-reject receiver exhibits the above I/Q mismatch values. Determine the image rejection ratio.

Solution:

The gain mismatch, $2(A_1 - A_2)/(A_1 + A_2) \approx (A_1 - A_2)/A_1 = \Delta A/A$, is obtained by raising 10 to the power of (0.2 dB/20) and subtracting 1 from the result. Thus,

$$\text{IRR} = \frac{4}{(\Delta A/A)^2 + \theta^2} \tag{13.8}$$

$$= 35\,\text{dB}. \tag{13.9}$$

In light of reported IRR values, the foregoing example suggests that this level of matching is possible without calibration. However, in practice it is difficult to maintain such stringent matching across the entire 11a band with a high yield. Most 11a/g receivers therefore employ I/Q calibration.

13.1.2 Transmitter

The transmitter chain must be linear enough to deliver a 64QAM OFDM signal to the antenna with acceptable distortion. In order to quantify the tolerable nonlinearity, a TX or PA model must be assumed and simulated with such a signal. The quality of the output is then expressed in terms of the bit error rate or the error vector magnitude. For example, the [5] employs the Rapp (static) model [4]:

$$g(V_{in}) = \frac{\alpha V_{in}}{[1 + (\frac{V_{in}}{V_0})^{2m}]^{\frac{1}{2m}}}, \tag{13.10}$$

where α denotes the small-signal gain around $V_{in} = 0$, and V_0 and m are fitting parameters. For typical CMOS PAs, $m \approx 2$ [5]. A 64QAM OFDM signal experiencing this nonlinearity yields the EVM shown in Fig. 13.4 as a function of the back-off from P_{1dB}. It is observed that a back-off of about 8 dB is necessary to meet the 11a/g specification, as also mentioned

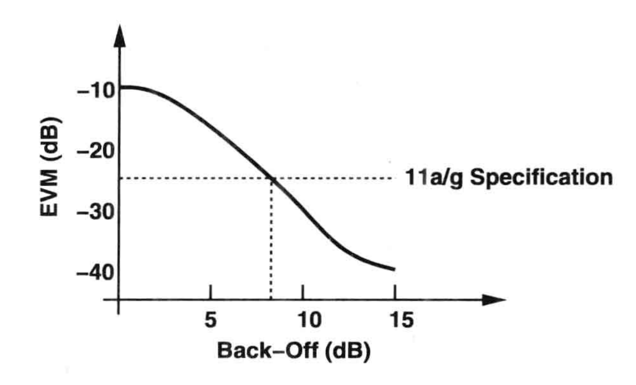


Figure 13.4 *EVM characteritics as a function of back-off.*

in [3]. Thus, for an output power of 40 mW (= +16 dBm), the TX output P_{1dB} must exceed approximately +24 dBm.[5]

As explained in Chapter 4, two TX design principles help achieve a high linearity: (1) assign most of the gain to the last PA stage so as to minimize the output swing of the preceding stages, and (2) minimize the number of stages in the TX chain.

Example 13.4

An 11a/g TX employs a two-stage PA having a gain of 15 dB. Can a quadrature upconverter directly drive this PA?

Solution:

The output P_{1dB} of the upconverter must exceed +24 dBm − 15 dB = +9 dBm = 1.78 V_{pp}. It is difficult to achieve such a high P_{1dB} at the output of typical mixers. A more practical approach therefore attempts to raise the PA gain or interposes another gain stage between the upconverter and the PA.

The gain of the TX chain from the baseband to the antenna somewhat depends on the design details. For example, a baseband swing of 0.2 V_{pp} requires a gain of 20 to reach an output swing of 4 V_{pp} (= +16 dBm).[6] As explained in Chapter 4, it is desirable to employ a relatively *large* baseband swing so as to minimize the effect of dc offsets and hence the carrier feedthrough, but mixer nonlinearity constrains this choice. For now, we assume a differential baseband swing of 0.2 V_{pp} in each of the I and Q paths.

The I/Q imbalance necessary in the TX is similar to that given for the RX in Section 13.1.1 (0.2 dB and 1.5°), requiring calibration in the transmit path as well. The

5. The simulations in [1] suggest a P_{1dB} of 20.5 dBm. The discrepancy possibly arises from different PA models.

6. With both I and Q inputs, the output voltage swing is higher by a factor of $\sqrt{2}$.

carrier feedthrough is another source of corruption in direct-conversion transmitters. For 11a/g systems, a feedthrough of about −40 dBc is achieved by means of baseband offset cancellation [5].

13.1.3 Frequency Synthesizer

For the dual-band transceiver developed in this chapter, the synthesizer must cover the 2.4-GHz and 5-GHz bands with a channel spacing of 20 MHz. In addition, the synthesizer must achieve acceptable phase noise and spur levels. We defer the band coverage issues to Section 13.1.4 and focus on the latter two here.

The phase noise in the receive mode creates reciprocal mixing and corrupts the signal constellation. The former effect must be quantified with the adjacent channels present. The following example illustrates the procedure.

Example 13.5

Determine the required synthesizer phase noise for an 11a receiver such that reciprocal mixing is negligible.

Solution:

We consider the high-sensitivity case, with the desired input at −82 dBm+3 dB and the adjacent and alternate channels at +16 dB and +32 dB, respectively. Figure 13.5 shows the corresponding spectrum but with the adjacent channels modeled as narrowband blockers to simplify the analysis. Upon mixing with the LO, the three components emerge in the baseband, with the phase noise skirts of the adjacent channels corrupting the desired signal. Since the synthesizer loop bandwidth is likely to be much smaller than 20 MHz, we can approximate the phase noise skirts by $S_\phi(f) = \alpha/f^2$.[7] Our objective is to determine α.

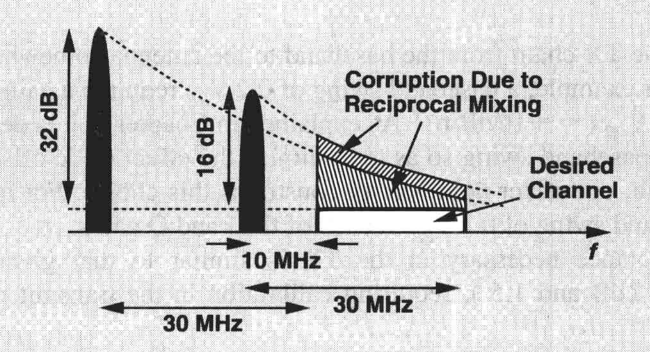


Figure 13.5 *Reciprocal mixing of two unequal blockers with a noisy LO.*

7. In this chapter, we represent the phase noise profile by either α/f^2 (assuming a center frequency of zero) or $\alpha/(f - f_c)^2$ (assuming a center frequency of f_c).

Example 13.5 *(Continued)*

If a blocker has a power that is a times the desired signal power, P_{sig}, then the phase noise power, P_{PN}, between frequency offsets of f_1 and f_2 and normalized to P_{sig} is given by

$$\frac{P_{PN}}{P_{sig}} = a \int_{f_1}^{f_2} \frac{\alpha}{f^2} df \tag{13.11}$$

$$= a\alpha \left(\frac{1}{f_1} - \frac{1}{f_2} \right). \tag{13.12}$$

In the scenario of Fig. 13.5, the total noise-to-signal ratio is equal to

$$\frac{P_{PN,tot}}{P_{sig}} = a_1\alpha \left(\frac{1}{f_1} - \frac{1}{f_2} \right) + a_2\alpha \left(\frac{1}{f_3} - \frac{1}{f_4} \right), \tag{13.13}$$

where $a_1 = 39.8$ ($= 16$ dB), $f_1 = 10$ MHz, $f_2 = 30$ MHz, $a_2 = 1585$ ($= 32$ dB), $f_3 = 30$ MHz, $f_4 = 50$ MHz. Note that the second term is much greater than the first in this case.

We wish to ensure that reciprocal mixing negligibly corrupts the signal; e.g., we target $P_{PN,tot}/P_{sig} = -20$ dB. It follows that $\alpha \approx 420$ and hence

$$S_n(f) = \frac{420}{f^2}. \tag{13.14}$$

For example, $S_n(f)$ is equal to -94 dBc/Hz at 1-MHz offset and -120 dBc/Hz at 20-MHz offset.

In the absence of reciprocal mixing, the synthesizer phase noise still corrupts the signal constellation. For this effect to be negligible in 11a/g, the total integrated phase noise must remain less than 1° [3]. To compute the integrated phase noise, P_ϕ, we approximate the synthesizer output spectrum as shown in Fig. 13.6: with a plateau from f_c to the edge of the

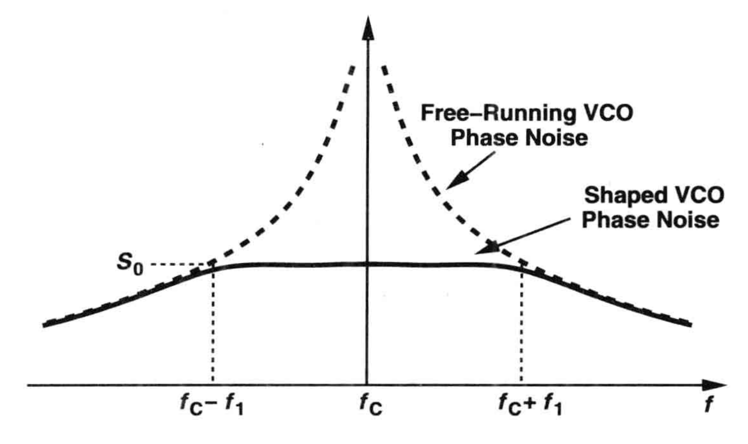


Figure 13.6 *Typical phase-locked phase noise profile.*

synthesizer loop bandwidth ($f_c \pm f_1$) and a declining profile given by $\alpha/(f-f_c)^2$ beyond $f_c \pm f_1$. Denoting the value of $\alpha/(f-f_c)^2$ at $f = f_c \pm f_1$ by S_0, we have $\alpha = S_0 f_1^2$ and

$$P_\phi = 2\int_{f_c}^{\infty} S_n(f)df \tag{13.15}$$

$$= 2S_0 f_1 + 2\int_{f_c+f_1}^{\infty} \frac{f_1^2 S_0}{(f-f_c)^2}df \tag{13.16}$$

$$= 2S_0 f_1 + 2S_0 f_1 \tag{13.17}$$

$$= 4S_0 f_1. \tag{13.18}$$

Let us assume that the synthesizer loop bandwidth, f_1, is about one-tenth of the channel spacing. For $\sqrt{P_\phi}$ to be less than $1^\circ = 0.0175$ rad, we have $S_0 = 3.83 \times 10^{-11}$ rad^2/Hz = -104 dBc/Hz. That is, the phase noise of the free-running VCO must be less than -104 dBc/Hz at 2-MHz offset, a more stringent requirement than the phase noise obtained in the above example at 1-MHz offset. The actual phase noise must be 3 dB lower to accommodate the TX VCO corruption as well. We will therefore bear in mind a target free-running phase noise of $-104 + 6 - 3 = -101$ dBc/Hz at 1-MHz offset.

Example 13.6

Having derived Eq. (13.18), a student reasons that a greater free-running phase noise, S_0, can be tolerated if the synthesizer loop bandwidth is *reduced*. Thus, f_1 must be minimized. Explain the flaw in this argument.

Solution:

Consider two scenarios with VCO phase noise profiles given by α_1/f^2 and α_2/f^2 (Fig. 13.7). Suppose the loop bandwidth is reduced from f_1 to $f_1/2$ and S_0 is allowed to rise to $2S_0$ so as to maintain P_ϕ constant. In the former case,

$$S_n(f_1) = \frac{\alpha_1}{f_1^2} = S_0 \tag{13.19}$$

and hence $\alpha_1 = f_1^2 S_0$. In the latter case,

$$S_n\left(\frac{f_1}{2}\right) = \frac{\alpha_2}{(0.5f_1)^2} = 2S_0, \tag{13.20}$$

and hence $\alpha_2 = 0.5 f_1^2 S_0$. It follows that the latter case demands a *lower* free-running phase noise at an offset of f_1, making the VCO design more difficult.

Example 13.6 *(Continued)*

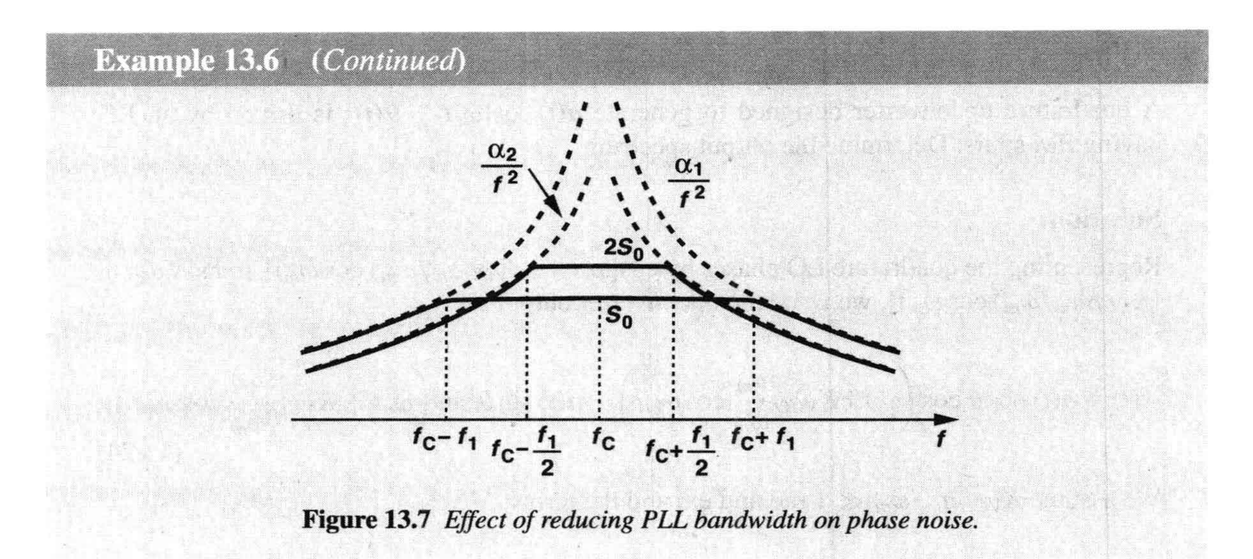


Figure 13.7 *Effect of reducing PLL bandwidth on phase noise.*

The synthesizer output spurs must also be considered. For an input level of −82 dBm +3 dB = −79 dBm, spurs in the middle of the adjacent and alternate channels downconvert blockers that are 16 dB and 32 dB higher, respectively. Thus, the spur levels at 20-MHz and 40-MHz offset must be below roughly −36 dBc and −52 dBc, respectively, so that each introduces a corruption of −20 dB. These specifications are relatively relaxed.

The spurs also impact the transmitted signal. To estimate the tolerable spur level, we return to the 1° phase error mentioned above (for random phase noise) and force the same requirement upon the effect of the (FM) spurs. Even though the latter are not random, we expect their effect on the EVM to be similar to that of phase noise. To this end, let us express the TX output in two cases, only with phase noise, $\phi_n(t)$:

$$x_{TX1}(t) = a(t)\cos[\omega_c t + \theta(t) + \phi_n(t)] \tag{13.21}$$

and only with a small FM spur

$$x_{TX2}(t) = a(t)\cos\left[\omega_c t + \theta(t) + K_{VCO}\frac{a_m}{\omega_m}\cos\omega_m t\right]. \tag{13.22}$$

For the total rms phase deviation to be less than $1^\circ = 0.0175$ rad, we have

$$\frac{K_{VCO}a_m}{\sqrt{2}\omega_m} = 0.0175. \tag{13.23}$$

The relative sideband level in $x_{TX2}(t)$ is equal to $K_{VCO}a_m/(2\omega_m) = 0.0124 = -38$ dBc.

Example 13.7

A quadrature upconverter designed to generate $a(t)\cos[\omega_c t + \theta(t)]$ is driven by an LO having FM spurs. Determine the output spectrum.

Solution:

Representing the quadrature LO phases by $\cos[\omega_c t + (K_{VCO}a_m/\omega_m)\cos\omega_m t]$ and $\sin[\omega_c t + (K_{VCO}a_m/\omega_m)\cos\omega_m t]$, we write the upconverter output as

$$x(t) = a(t)\cos\theta\cos\left(\omega_c t + K_{VCO}\frac{a_m}{\omega_m}\cos\omega_m t\right) - a(t)\sin\theta\sin\left(\omega_c t + K_{VCO}\frac{a_m}{\omega_m}\cos\omega_m t\right). \tag{13.24}$$

We assume $K_{VCO}a_m/\omega_m \ll 1$ rad and expand the terms:

$$x(t) \approx a(t)\cos\theta\cos\omega_c t - a(t)\sin\theta\sin\omega_c t - K_{VCO}\frac{a_m}{\omega_m}\cos\omega_m t a(t)\cos\theta\sin\omega_c t$$

$$- K_{VCO}\frac{a_m}{\omega_m}\cos\omega_m t a(t)\sin\theta\cos\omega_c t \tag{13.25}$$

$$\approx a(t)\cos(\omega_c t + \theta) - K_{VCO}\frac{a_m}{\omega_m}\cos\omega_m t a(t)\sin(\omega_c t + \theta). \tag{13.26}$$

The output thus contains the desirable component and the quadrature of the desirable component shifted to center frequencies of $\omega_c - \omega_m$ and $\omega_c + \omega_m$ (Fig. 13.8). The key point here is that the synthesizer spurs are modulated as they emerge in the TX path.

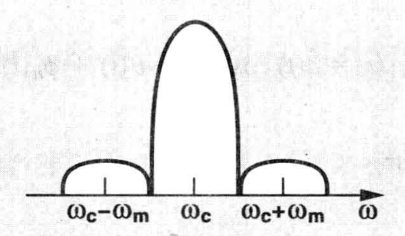


Figure 13.8 *Modulation of synthesizer spurs in a transmitter.*

13.1.4 Frequency Planning

A direct-conversion transceiver is a natural choice for our 11a/g system. However, it is not obvious how the necessary LO frequencies and phases should be generated. We wish to cover approximately 5.1 GHz to 5.9 GHz for 11a and 2.400 GHz to 2.480 GHz for 11g while providing quadrature outputs and avoiding LO pulling in the TX mode.

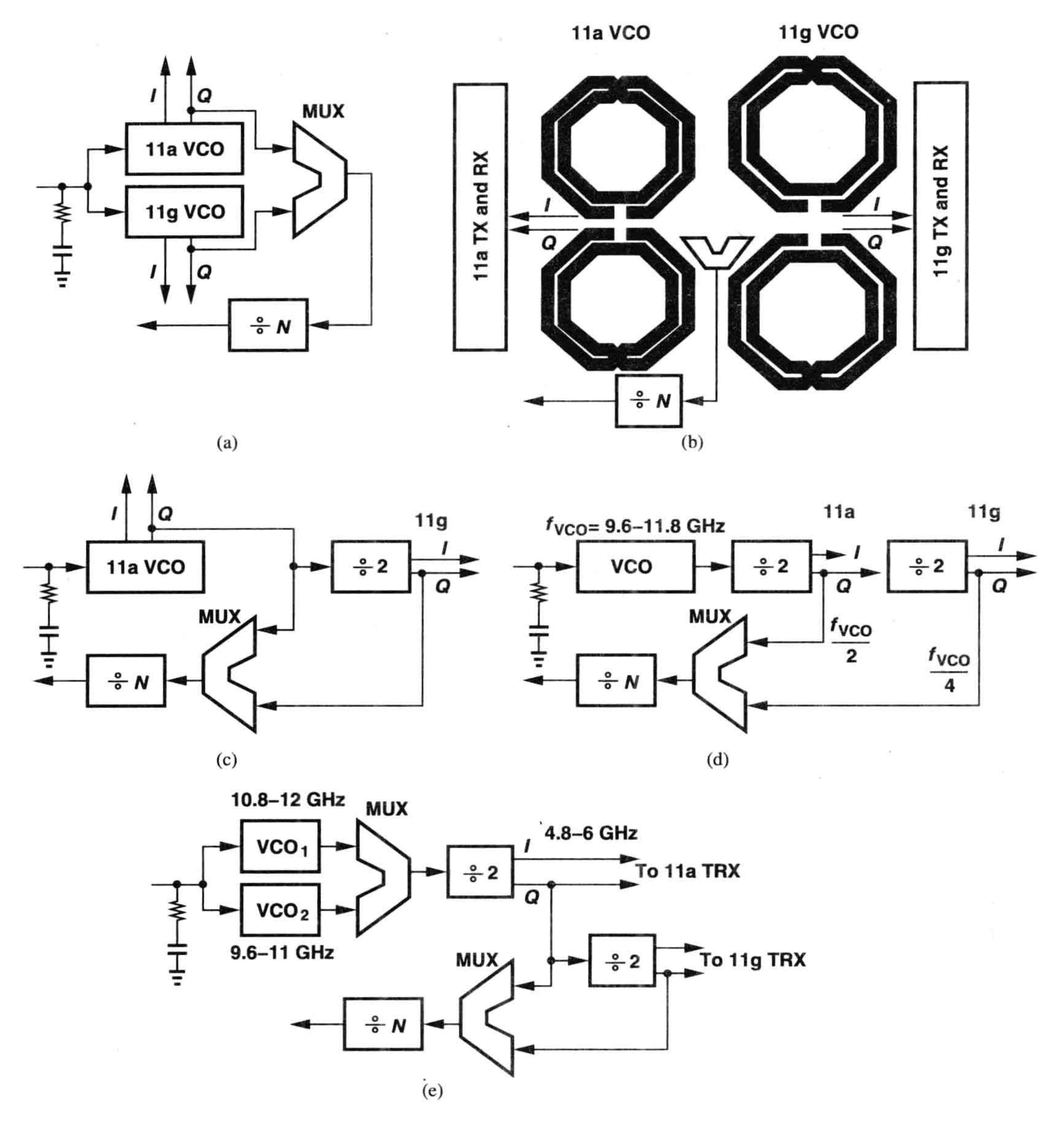


Figure 13.9 *(a) Use of two VCOs for 11a and 11g bands, (b) TRX floor plan for (a), (c) use of a VCO and a divider for the two bands, (d) use of a VCO at twice the carrier frequency to avoid injection pulling, and (e) use of two VCOs to relax tuning range requirement.*

Let us consider several different approaches.

- Two separate quadrature VCOs for the two bands, with their outputs multiplexed and applied to the feedback divider chain [Fig. 13.9(a)]. In this case, the four VCO inductors lead to the floor plan shown in Fig. 13.9(b), imposing a large spacing between the 11a and 11g signal paths. This issue becomes critical if the two

paths are to share high-frequency circuits (e.g., LNAs and mixers). Also, the 11a VCO must provide a tuning range of about ±15%. Finally, LO pulling proves serious.

- One quadrature VCO serving both bands [Fig. 13.9(c)]. Here, the floor plan is more compact, but the VCO must tune from 4.8 GHz to 5.9 GHz, i.e., by about ±21%. The issue of LO pulling persists for the 11a band and is somewhat serious for the 11g band if the second harmonic of the 11g PA output couples to the VCO. For this reason, it is desirable to implement the 11g PA in fully-differential form [but without symmetric inductors (Chapter 7)].
- One differential VCO operating from 2×4.8 GHz to 2×5.9 GHz [Fig. 13.9(d)]. This choice allows a compact floor plan but requires (1) a tuning range of ±21%, (2) differential 11a and 11g PAs, and (3) a $\div 2$ circuit that robustly operates up to 12 GHz, preferably with no inductors. Fortunately, the raw speed of transistors in 65-nm CMOS technology permits such a divider design.

Example 13.8

Explain why the outputs of the two $\div 2$ circuits in Fig. 13.9(d) are multiplexed. That is, why do we not apply the $f_{VCO}/4$ output to the $\div N$ stage in the 11a mode as well?

Solution:

Driving the $\div N$ stage by $f_{VCO}/4$ is indeed desirable as it eases the design of this circuit. However, in an integer-N architecture, this choice calls for a reference frequency of 10 MHz rather than 20 MHz in the 11a mode (why?), leading to a smaller loop bandwidth and less suppression of the VCO phase noise. In other words, *if* the VCO provides sufficiently low phase noise, *then* the $\div N$ stage can be driven by $f_{VCO}/4$ in both modes.

We expect that the relatively high operation frequency and wide tuning range required of the VCO in Fig. 13.9(d) inevitably result in a high phase noise. We therefore employ two VCOs, each with about half the tuning range but with some overlap to avoid a blind zone [Fig. 13.9(e)]. A larger number of VCOs can be utilized to allow an even narrower tuning range for each, but the necessary additional inductors complicate the routing.

Example 13.9

The MUX following the two VCOs in Fig. 13.9(e) must either consume a high power or employ inductors. Is it possible to follow each VCO by a $\div 2$ circuit and perform the multiplexing at the dividers' outputs?

Example 13.9 (*Continued*)

Solution:

Illustrated in Fig. 13.10, this approach is indeed superior (if the ÷2 circuits do not need inductors). The two multiplexers do introduce additional I/Q mismatch, but calibration removes this error along with other blocks' contributions. Note that the new ÷2 circuit does not raise the power consumption because it is turned off along with VCO_2 when not needed.

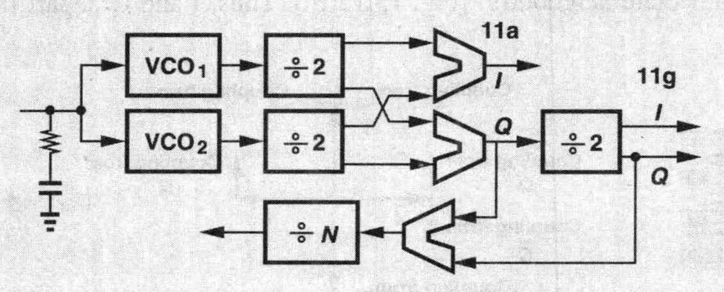


Figure 13.10 *Use of MUXes after dividers.*

The frequency plan depicted in Fig. 13.9(e) resolves most of the issues that we have encountered, with the proviso that the two PAs are implemented differentially. We must now decide how the synthesizer is shared between the TX and RX paths. Shown in Fig. 13.11 is one scenario where the synthesizer outputs directly drive both paths. In practice, buffers may be necessary before and after the long wires.

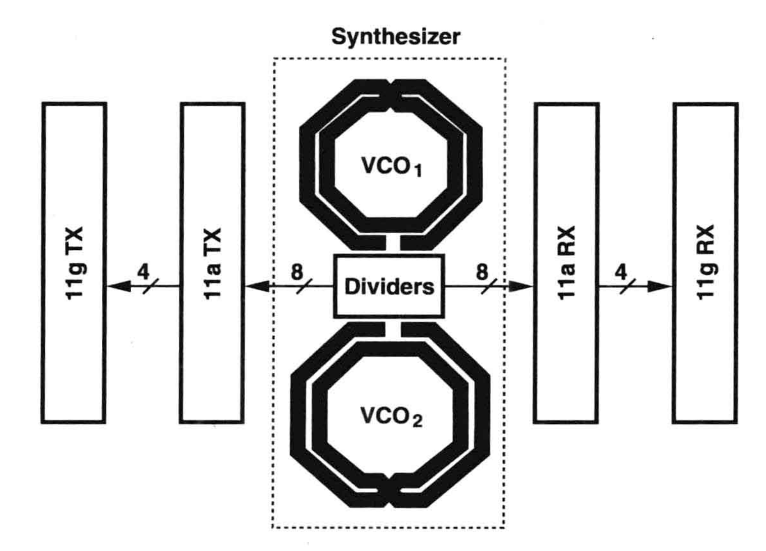


Figure 13.11 *TRX floor plan with two VCOs running at twice the carrier frequency.*

Example 13.10

Differential I and Q signals experience deterministic mismatches as they travel on long interconnects. Explain why and devise a method of suppressing this effect.

Solution:

Consider the arrangement shown in Fig. 13.12(a). Owing to the finite resistance and coupling capacitance of the wires, each line experiences an additive fraction of the signal(s) on its immediate neighbor(s) [Fig. 13.12(b)]. Thus, I and $\overline{Q}$ depart from their ideal orientations.

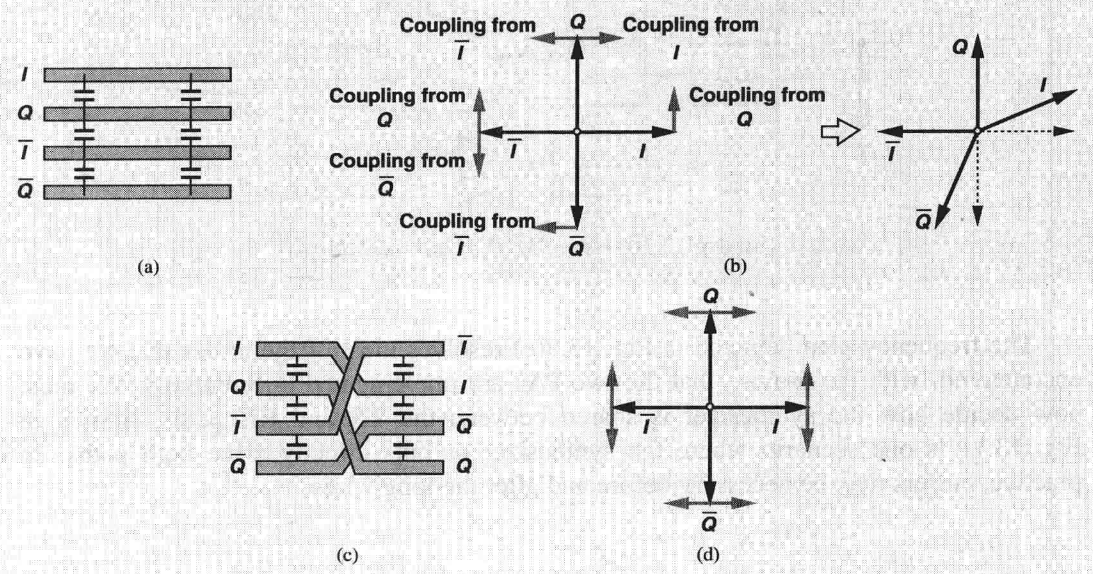


Figure 13.12 *(a) Lines carrying I and Q LO phases, (b) mismatches resulting from coupling, (c) cross routing scheme, (d) cancellation of mismtaches.*

To suppress this effect, we rearrange the wires as shown in Fig. 13.12(c) at half of the distance between the end points, creating a different set of couplings. Illustrated in Fig. 13.12(d) are all of the couplings among the wires, revealing complete cancellation.

Figure 13.13 shows the overall transceiver architecture developed so far. As seen later, the same RX path can in fact be used for 11a and 11g.

13.2 RECEIVER DESIGN

The 11a/g receiver chains are designed for their respective input frequency ranges with the required NF, linearity, gain, and automatic gain control (AGC). The AGC is realized by discrete gain control along the chain and controlled by the digital inputs provided by the baseband processor.

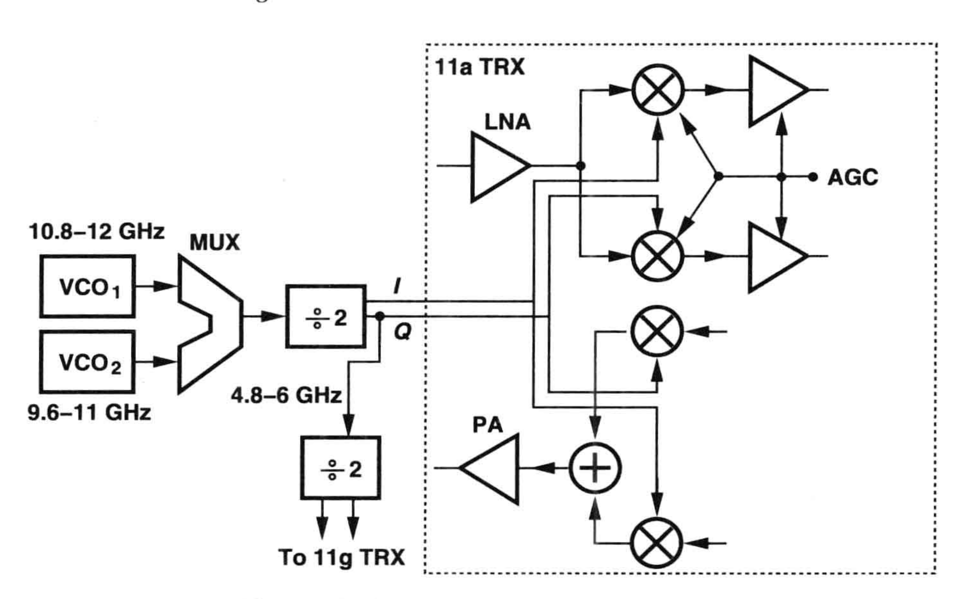


Figure 13.13 *Final transceiver architecture.*

13.2.1 LNA Design

The two 5-GHz design examples described in Chapter 5 are candidates for the 11a receiver. But is it possible to employ only *one* LNA for the two bands? Let us explore another LNA topology here.

Consider the resistive-feedback LNA shown in Fig. 13.14(a). Here, M_2 serves as both a load and an amplifying device, yielding a lower noise figure than if the load is passive. Current source I_1 defines the bias of M_1 and M_2, and C_1 creates an ac ground at node X. This circuit can potentially cover the frequency range of 2.4 GHz to 6 GHz. In Problem 13.7, we prove that

$$\frac{V_{out}}{V_{in}} = -\frac{[1-(g_{m1}+g_{m2})R_F](r_{O1}||r_{O2})}{R_F+R_S+[1+(g_{m1}+g_{m2})R_S](r_{O1}||r_{O2})}, \tag{13.27}$$

and

$$R_{in} = \frac{r_{O1}||r_{O2}+R_F}{1+(g_{m1}+g_{m2})(r_{O1}||r_{O2})}. \tag{13.28}$$

Equating R_{in} to R_S and making a substitution in the denominator of Eq. (13.27), we have

$$\frac{V_{out}}{V_{in}} = -\frac{[1-(g_{m1}+g_{m2})R_F](r_{O1}||r_{O2})}{2(R_F+r_{O1}||r_{O2})}. \tag{13.29}$$

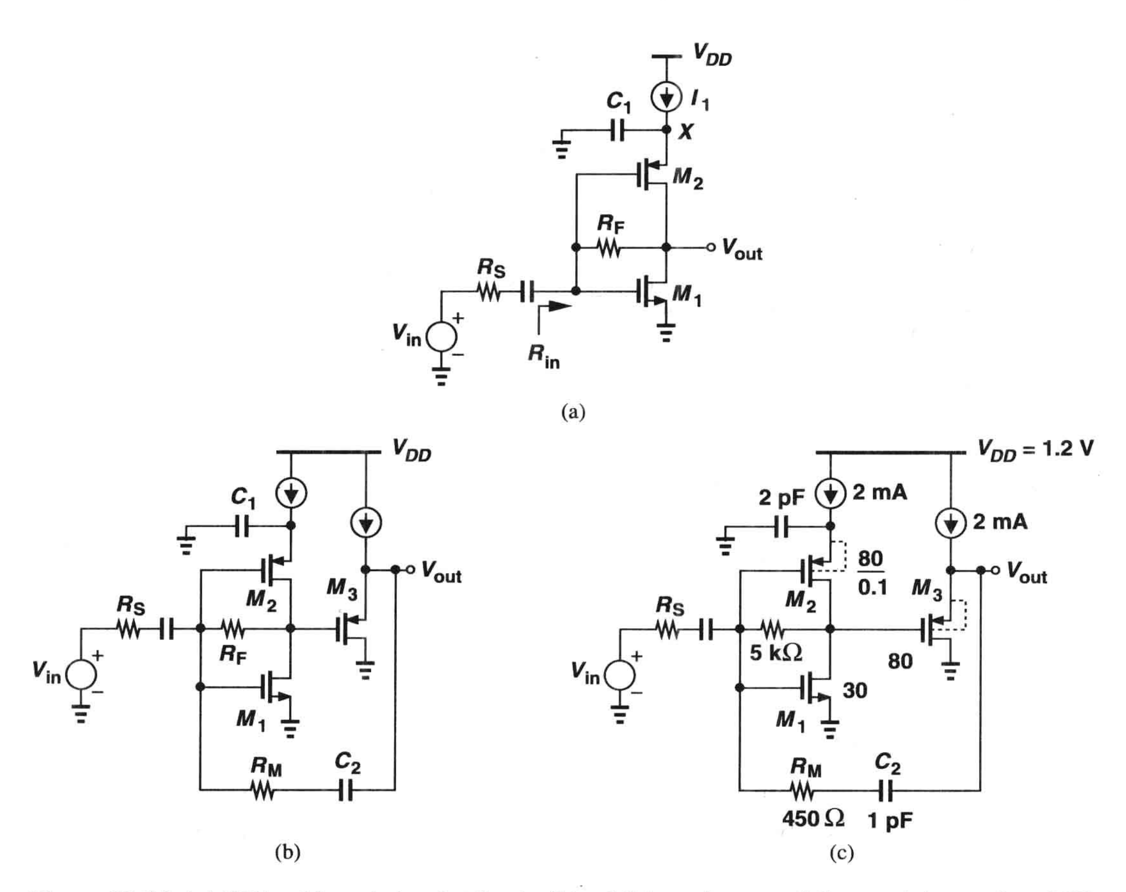


Figure 13.14 *(a) LNA with resistive feedback, (b) addition of source follower, (c) complete LNA design.*

We now make two observations based on rough estimates. Suppose $(g_{m1} + g_{m2})(r_{O1}||r_{O2}) \gg 1$. First, from Eq. (13.28),

$$R_{in} \approx \frac{1}{g_{m1} + g_{m2}} + \frac{R_F}{(g_{m1} + g_{m2})(r_{O1}||r_{O2})}. \tag{13.30}$$

For $R_{in} \approx 50\ \Omega$, we surmise that the first term should be on the order of 10 to 20 Ω (as it affects the noise figure) and the second, 30 to 40 Ω. That is, R_F cannot exceed 300 to 400 Ω if $(g_{m1} + g_{m2})(r_{O1}||r_{O2})$ is around 10. Second, from Eq. (13.29), we can compute the gain with $g_{m1}+g_{m2} \approx (20\ \Omega)^{-1}$, $r_{O1}||r_{O2} \approx 200\ \Omega$,[8] and $R_F = 300\ \Omega$, obtaining $V_{out}/V_{in} = -2.8$. In practice, minimum-length devices in 65-nm technology yield a smaller value for $(g_{m1} + g_{m2})(r_{O1}||r_{O2})$ and hence even a lower gain. The circuit thus suffers from a tight trade-off between the input matching and the gain.

In order to achieve a higher gain while providing input matching, we modify the circuit to that shown in Fig. 13.14(b). In this case, R_F is large, only establishing proper dc level

8. Since $g_m r_O \approx 10$ and $g_m \approx (40\ \Omega)^{-1}$, we have $r_O \approx 400\ \Omega$.

at the gates of M_1 and M_2 and allowing a higher voltage gain. The source follower, on the other hand, drives a moderate resistance, R_M, to match the input. For a large R_F and negligible body effect and channel-length modulation in M_3, the input resistance is given by the feedback resistance divided by one plus the loop gain:

$$R_{in} \approx \frac{R_M + g_{m3}^{-1}}{1 + (g_{m1} + g_{m2})(r_{O1} || r_{O2})}. \tag{13.31}$$

(Why is g_{m3}^{-1} included in the numerator?) Lacking the $r_{O1}||r_{O2}$ term in the numerator of (13.28), this result is more favorable as it permits a larger R_M. If $R_{in} = R_S$ and $R_M \gg g_{m3}^{-1}$, then the gain is simply equal to 1/2 times the voltage gain of the inverter:

$$\frac{V_{out}}{V_{in}} = -\frac{1}{2}(g_{m1} + g_{m2})(r_{O1}||r_{O2}). \tag{13.32}$$

For example, if $(g_{m1} + g_{m2})(r_{O1}||r_{O2}) = 10$, then a gain of 14 dB is obtained.

Figure 13.14(c) depicts the final LNA design. We should make a few remarks here. First, with a 1.2-V supply, $|V_{GS2}| + V_{GS1}$ must remain below about 1 V, requiring wide transistors. Second, to increase the gain, the channel length of M_2 is raised to 0.1 μm. Third, to minimize $|V_{GS2}|$ and $|V_{GS3}|$, the n-well of each device is tied to its source.

Example 13.11

The large input transistors in Fig. 13.14(c) present an input capacitance, C_{in}, of about 200 fF (including the Miller effect of $C_{GD1}+C_{GD2}$). Does this capacitance not degrade the input match at 6 GHz?

Solution:

Since $(C_{in}\omega)^{-1} \approx 130\ \Omega$ is comparable with 50 Ω, we expect C_{in} to affect S_{11} considerably. Fortunately, however, the capacitance at the *output* node of the inverter creates a pole that drops the open-loop gain at high frequencies, thus *raising* the closed-loop input impedance. This is another example of reactance-cancelling LNAs described in Chapter 5.

Figure 13.15 plots the simulated characteristics of the LNA across a frequency range of 2 GHz to 6 GHz. The worst-case $|S_{11}|$, NF, and gain[9] are equal to −16.5 dB, 2.35 dB, and 14.9 dB, respectively. Shown in Fig. 13.16 is the LNA gain as a function of the input level at 6 GHz. By virtue of negative feedback, the LNA achieves a P_{1dB} of about −14 dBm.[10]

13.2.2 Mixer Design

The choice between passive and active mixers depends on several factors, including available LO swings, required linearity, and output flicker noise. In this transceiver design,

9. This voltage gain is from the LNA input node to the output.

10. Note that ac and transient simulations yield slightly different voltage gains.

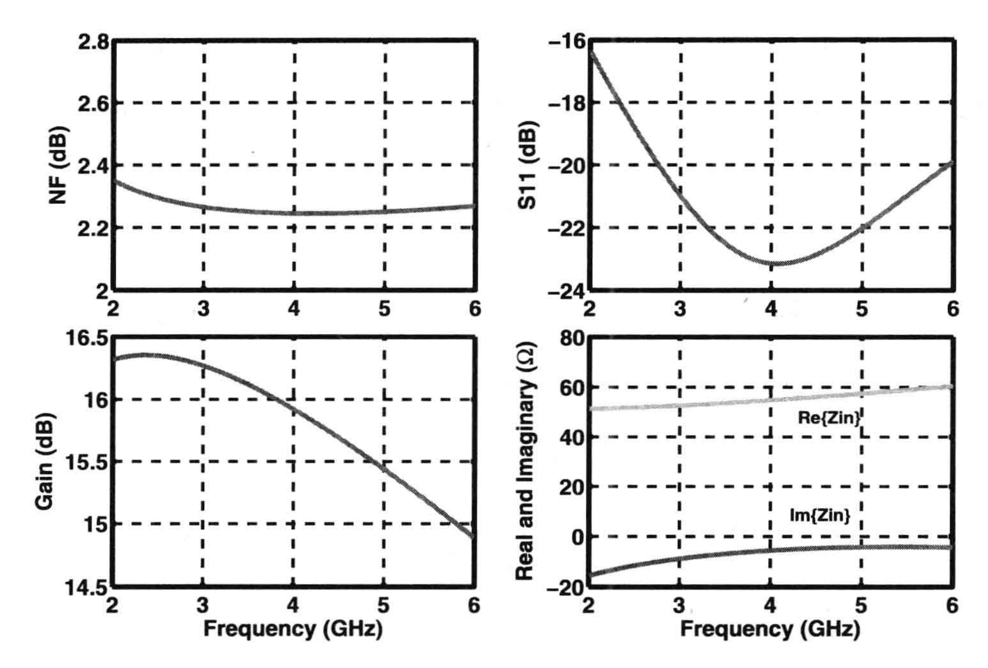


Figure 13.15 *Simulated characteristics of 11a/g LNA.*

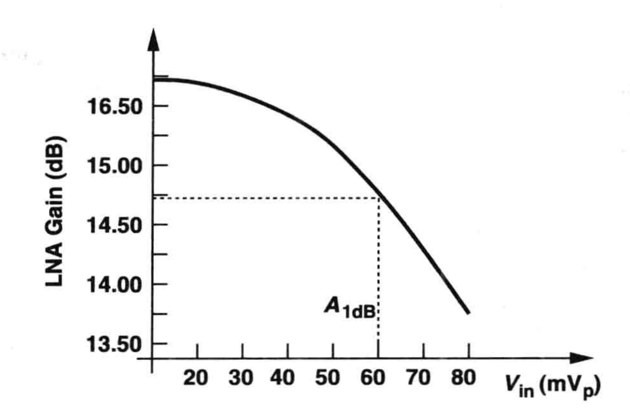


Figure 13.16 *LNA compression characteristic.*

we have some flexibility because (a) 65-nm CMOS technology can provide rail-to-rail LO swings at 6 GHz, allowing passive mixers, and (b) the RX linearity is relatively relaxed, allowing active mixers. Nonetheless, the high flicker noise of 65-nm devices proves problematic in active topologies.

We consider a single-balanced passive mixer followed by a simple baseband amplifier (Fig. 13.17). Here, to minimize the amplifier's flicker noise, large PMOS devices are employed. The gate bias voltage of the differential pair is defined by V_b and is 0.2 V above

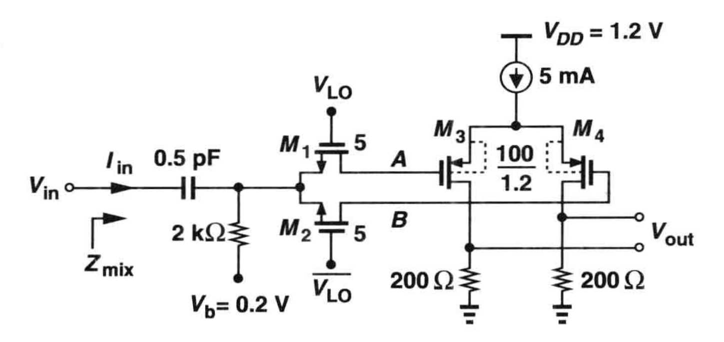


Figure 13.17 *Downconversion mixer design.*

ground to ensure M_3 and M_4 operate in saturation. Note that two instances of this chain are required for quadrature downconversion, drawing a total supply current of 10 mA.

Using the equations derived for voltage-driven sampling (non-return-to-zero) mixers in Chapter 6, we can compute the characteristics of the above circuit. Transistors M_3 and M_4 present a load capacitance of $C_L \approx (2/3)WLC_{ox} \approx 130$ fF to the mixer devices. The differential noise measured between A and B is thus given by

$$\overline{V_{n,AB}^2} = 2kT\left(3.9R_{1,2} + \frac{1}{2C_L f_{LO}}\right), \tag{13.33}$$

where $R_{1,2}$ denotes the on-resistance of M_1 and M_2 and is about 100 Ω. It follows that $\overline{V_{n,AB}^2} \approx 8.54 \times 10^{-18}$ V^2. Assuming a voltage gain of about unity from V_{in} to V_{AB}, we can determine the noise figure with respect to a 50-Ω source by dividing $\overline{V_{n,AB}^2}$ by the noise of a 50-Ω resistor[11] and adding 1 to the result. That is, NF = 11.31 = 10.1 dB at f_{LO} = 6 GHz. Simulations confirm this value and reveal negligible flicker noise at A and B.

The circuit of Fig. 13.17 entails a number of issues. First, though incorporating large transistors, the differential pair still contributes significant flicker noise, raising the NF by several dB at 100 kHz. The trade-off here lies between the impedance that this chain presents to the LNA and the flicker noise of M_3 and M_4.

Second, the LNA must drive *four* switches and their sampling capacitors, thereby sustaining a heavy load. Thus, the LNA gain and input matching may degrade. In other words, the LNA and mixer designs must be optimized as one entity.

Third, the inverse dependence of $V_{n,AB}$ upon f_{LO} in Eq. (13.33) implies that the mixer suffers from a *higher* noise figure in the 11g band. This is partially compensated by the higher input impedance of the mixer and hence greater LNA gain.

Figure 13.18 plots the simulated double-sideband noise figure of the mixer of Fig. 13.17 with respect to a 50-Ω source impedance. For a 6-GHz LO, the NF is dominated by the flicker noise of the baseband amplifier at 100-kHz offset. For a 2.4-GHz LO, the thermal noise floor rises by 3 dB. The simulations assume a rail-to-rail sinusoidal LO waveform.

11. We assume that the input impedance of the mixer is much higher than 50 Ω.

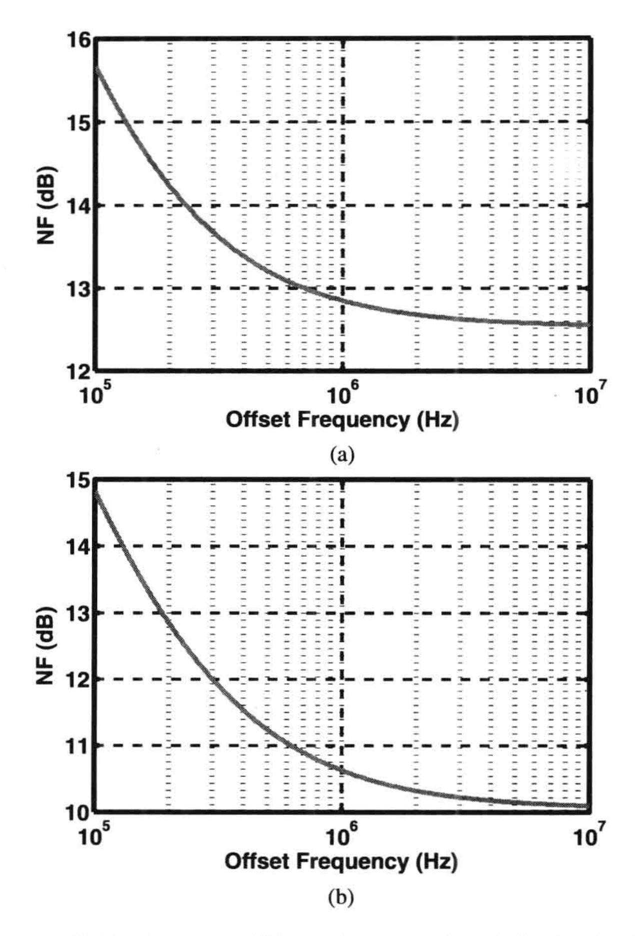


Figure 13.18 *Simulated NF of mixer at (a) 2.4 GHz, (b) 6 GHz.*

Figure 13.19 shows the overall RX chain, and Fig. 13.20 plots its simulated double-sideband noise figure. The RX noise figure varies from 7.5 dB to 6.1 dB at 2.4 GHz and from 7 dB to 4.5 dB at 6 GHz. These values are well within our target of 10 dB.

Example 13.12

How is the receiver sensitivity calculated if the noise figure varies with the frequency?

Solution:

A simple method is to translate the NF plot to an output noise spectral density plot and compute the total output noise power in the channel bandwidth (10 MHz). In such an approach, the flicker noise depicted in Fig. 13.20 contributes only slightly because most of its energy is carried between 100 kHz and 1 MHz.

In an OFDM system, on the other hand, the flicker noise corrupts some subchannels to a much greater extent than other subchannels. Thus, system simulations with the actual noise spectrum may be necessary.

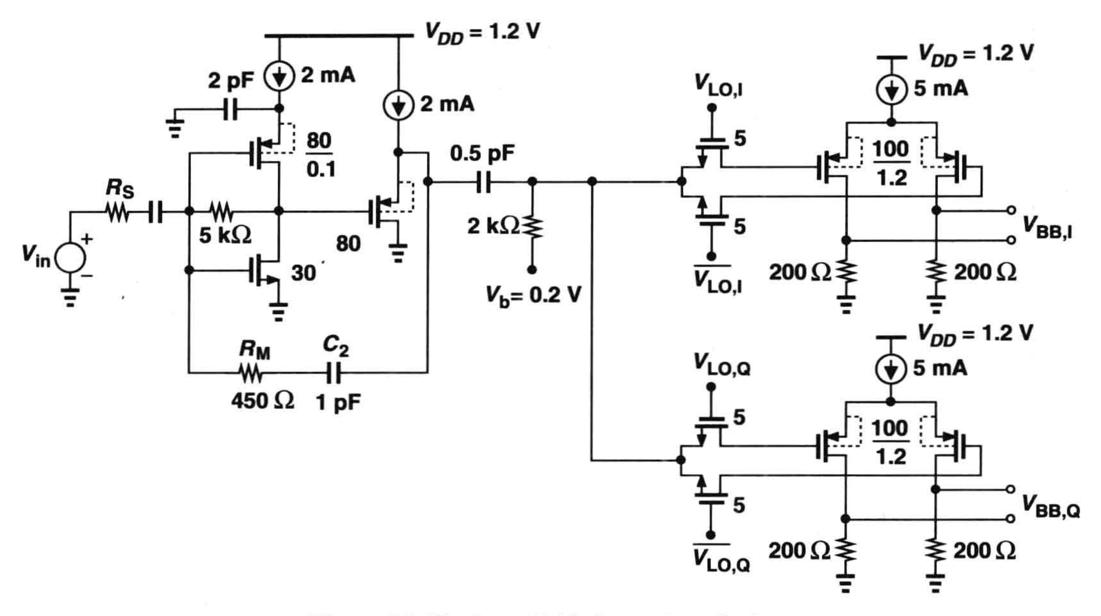


Figure 13.19 *Overall 11a/g receiver design.*

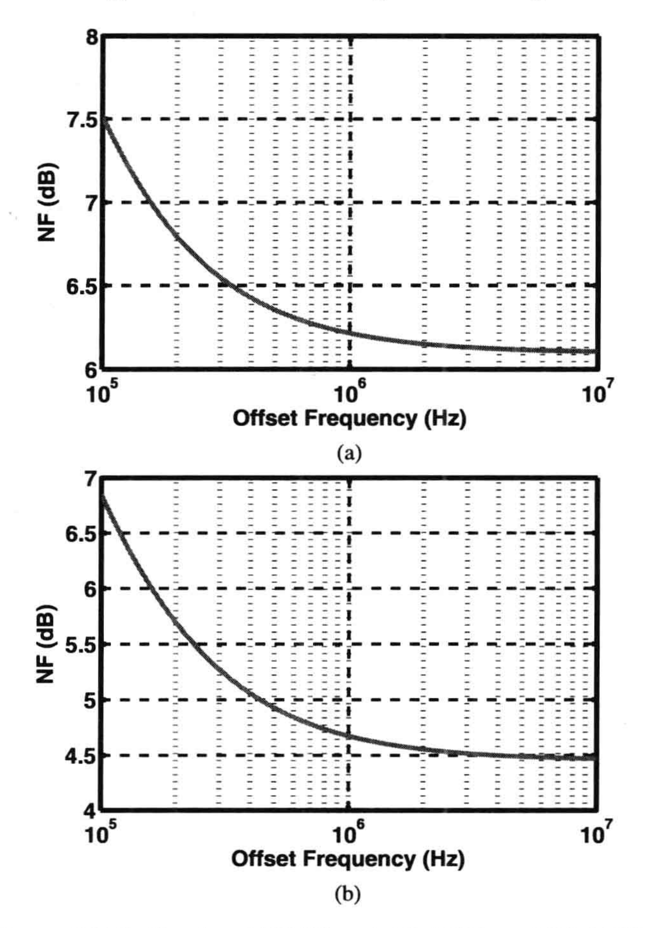


Figure 13.20 *Simulated RX NF at (a) 2.4 GHz and (b) 6 GHz.*

Example 13.13

The input impedance, Z_{mix}, in Fig. 13.17 may alter the feedback LNA input return loss. How is this effect quantified?

Solution:

The LNA S_{11} plot in Fig. 13.15 is obtained using small-signal ac simulations. On the other hand, the input impedance of passive mixers must be determined with the transistors switching, i.e., using transient simulations. To study the LNA input impedance while the mixers are switched, the FFT of I_{in} in Fig. 13.17 can be taken and its magnitude and phase plotted. With the amplitude and phase of V_{in} known, the input impedance can be calculated at the frequency of interest.

13.2.3 AGC

As mentioned in Section 13.1.1, the RX gain must be programmable from 23 dB to 58 dB so as to withstand a maximum input level of −30 dBm. The principal challenge in realizing a variable gain in the front end is to avoid altering the RX input impedance. For example, if resistor R_M in Fig. 13.14(c) varies, so does the S_{11}. Fortunately, as shown in Fig. 13.16, the LNA 1-dB compression point is well above −30 dBm, allowing a fixed LNA gain for the entire input level range.

In order to determine where in the receiver chain we must vary the gain, we first plot the overall RX gain characteristic (Figure 13.21), obtaining an input P_{1dB} of −26 dBm. Dominated by the baseband differential pair, the RX P_{1dB} is quite lower than that of the LNA. It is therefore desirable to lower the mixer gain as the average RX input level approaches −30 dBm, especially because the peak-to-average ratio of 11a/g signals can reach 9 dB. As shown in Fig. 13.22, this is accomplished by inserting transistors M_{G1}-M_{G3} between the differential outputs of the mixer. For an input level of around −50 dBm, M_{G1} is turned on, reducing the gain by about 5 dB. For an input level of −40 dBm, both M_{G1} and M_{G2} are turned on, lowering the gain by 10 dB. Finally, for an input level of −30 dBm, all three transistors are turned on, dropping the gain by 15 dB. Of course, we hope that the RX P_{1dB}

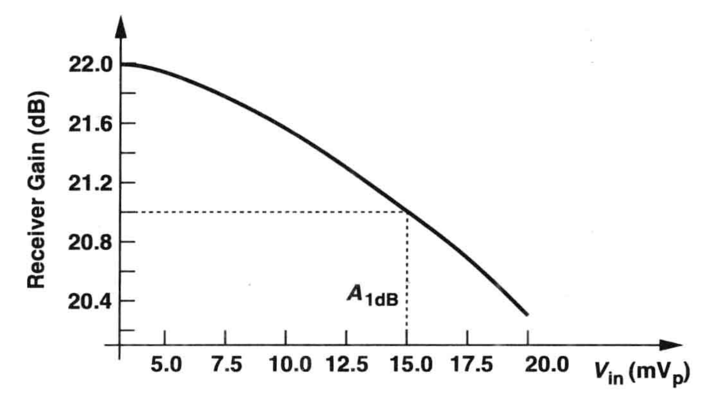


Figure 13.21 *Compression characteristic of 11a/g receiver.*

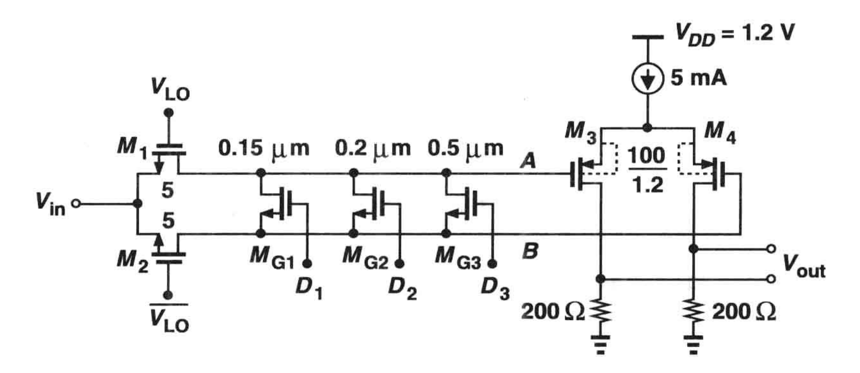


Figure 13.22 *Coarse AGC embedded within downconversion mixer.*

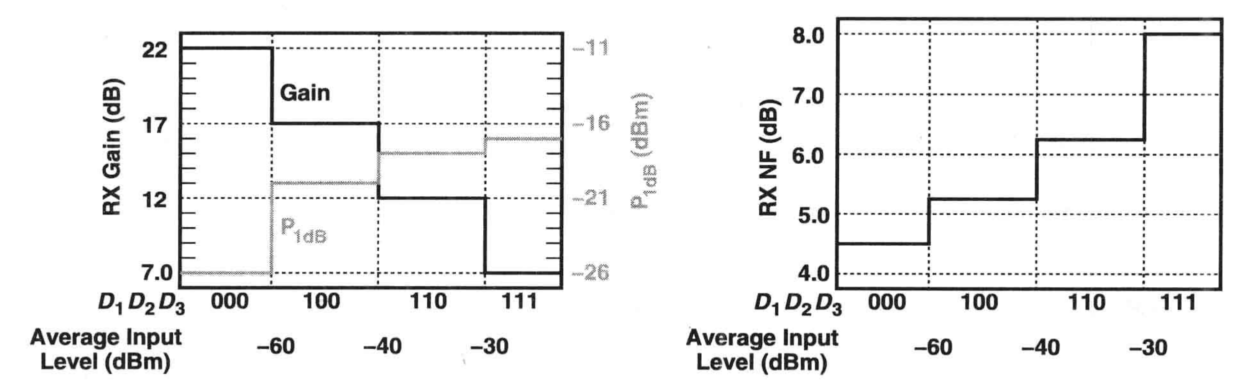


Figure 13.23 *Receiver performance as a function of gain setting.*

rises by approximately the same amount in each case, reaching a comfortable value in the low-gain mode. We call this arrangement the "coarse AGC."

The necessary widths of M_{G1}-M_{G3} are obtained from simulations to be 0.15 μm, 0.2 μm, and 0.5 μm, respectively ($L = 60$ nm). Figure 13.23 plots the receiver gain, P_{1dB}, and NF for the different gain settings.

We should make two remarks. First, owing to their small dimensions, M_{G1}-M_{G3} suffer from large threshold variations. It is therefore preferable to increase both the width and length of each device by a factor of 2 to 5 while maintaining the desired on-resistance. Second, the characteristics of Fig. 13.23 indicate that the RX P_{1dB} hardly exceeds -18 dBm even as the gain is lowered further. This is because, beyond this point, the nonlinearity of the LNA and mixer (rather than the baseband amplifier) dominates.

Example 13.14

What controls D_1–D_3 in Fig. 13.22?

Solution:

The digital control for D_1–D_3 is typically generated by the baseband processor. Measuring the signal level digitized by the baseband ADC, the processor determines how much attenuation is necessary.

With a maximum gain of 22 dB provided by the front end, the RX must realize roughly another 40 dB of gain in the baseband (Example 13.2) (the "fine AGC"). In practice, the amplification and channel-selection filtering are interspersed, thus relaxing the linearity of the gain stages.[12]

Example 13.15

What gain steps are required for the fine AGC?

Solution:

The fine gain step size trades with the baseband ADC resolution. To understand this point, consider the example shown in Fig. 13.24(a), where the gain changes by h dB for every 10-dB change in the input level. Thus, as the input level goes from, say, -39.9 dBm to -30.1 dBm, the gain is constant and hence the ADC input rises by 10 dB. The ADC must therefore (a) digitize the signal with proper resolution when the input is around -39.9 dBm, *and* (b) accommodate the signal without clipping when the input is around -30.1 dBm. In other words, the ADC must provide an additional 10 dB of dynamic range to avoid clipping its input as the received signal goes from -39.9 dBm to -30.1 dBm.

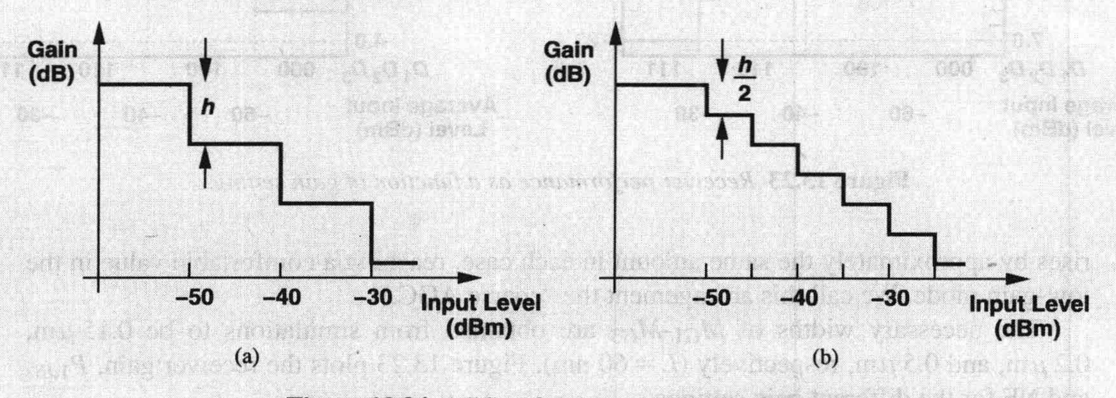


Figure 13.24 *AGC with (a) coarse and (b) fine steps.*

Now consider the scenario depicted in Fig. 13.24(b), where gain switching occurs for every 5-dB change in the input level. In this case, the ADC must provide 5 dB of additional resolution (dynamic range).

In order to minimize the burden on the ADC, the AGC typically employs a gain step of 1 or 2 dB. Of course, in systems with a narrow channel bandwidth, e.g., GSM, the baseband ADC runs at a relatively low speed and can be designed for a wide dynamic range, thereby relaxing the AGC requirements.

Another issue related to AGC is the variation of the baseband DC offset as the gain changes. Since switching the LNA or mixer gain may alter the amount of the LO coupling

12. The linearity is relaxed for the intermodulation of blockers but not for the compression of the desired signal.

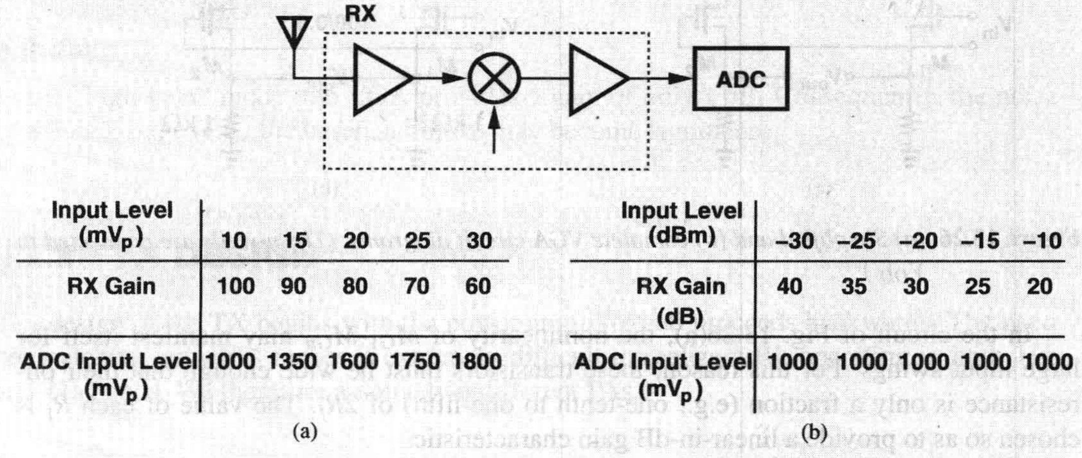
to the RX input and hence the self-mixing result, the DC offset changes. To deal with this effect, one can (1) perform offset cancellation for each gain setting and store the results in the digital domain so that the offset is corrected as the gain is switched, or (2) increase the ADC dynamic range to accommodate the uncorrected offset.

Example 13.16

In AGC design, we seek a programmable gain that is "linear in dB," i.e., for each LSB increase in the digital control, the gain changes by h dB and h is constant. Explain why.

Solution:

The baseband ADC and digital processor measure the signal amplitude and adjust the digital gain control. Let us consider two scenarios for the gain adjustment as a function of the signal level. As shown in Fig. 13.25(a), in the first scenario the (numerical) gain is reduced by *a constant (numerical) amount* (10) for a constant increase in the input amplitude (5 mV). In this case, the voltage swing sensed by the ADC (= input level × RX gain) is not constant, requiring nearly doubling the ADC dynamic range as the input varies from 10 mV_p to 30 mV_p.



Input Level (mV_p)	10	15	20	25	30
RX Gain	100	90	80	70	60
ADC Input Level (mV_p)	1000	1350	1600	1750	1800

(a)

Input Level (dBm)	−30	−25	−20	−15	−10
RX Gain (dB)	40	35	30	25	20
ADC Input Level (mV_p)	1000	1000	1000	1000	1000

(b)

Figure 13.25 *AGC with (a) linear and (b) logarithmic gain steps as a function of the input level.*

In the second scenario [Fig. 13.25(b)], the RX gain is reduced by a constant amount in dB for a constant logarithmic increase in the signal level, thereby keeping the ADC input swing constant. Here, for every 5 dB rise in the RX input, the baseband processor changes the digital control by 1 LSB, lowering the gain by 5 dB. It is therefore necessary to realize a linear-in-dB gain control mechanism, as accomplished in Fig. 13.23.

The baseband gain and filtering stages should negligibly degrade the RX noise and linearity. In practice, however, noise-linearity-power trade-offs make it difficult to fulfill this wish with a reasonable power consumption. Consequently, the linearity of typical

receivers (in the high-gain mode) is limited by that of the baseband stages rather than the front end.

We now implement the fine AGC. Figure 13.26(a) depicts a variable-gain amplifier (VGA)[13] suited for use in the baseband. Here, the gain is reduced by raising the degeneration resistance: in the high-gain mode, M_{G1}-M_{Gn} are on, and to lower the gain, we turn off M_{G1}; or M_{G1} and M_{G2}; or M_{G1}, M_{G2}, and M_{G3}; etc. Note that as the gain falls, the stage becomes more linear, a desirable and even necessary behavior for VGAs.

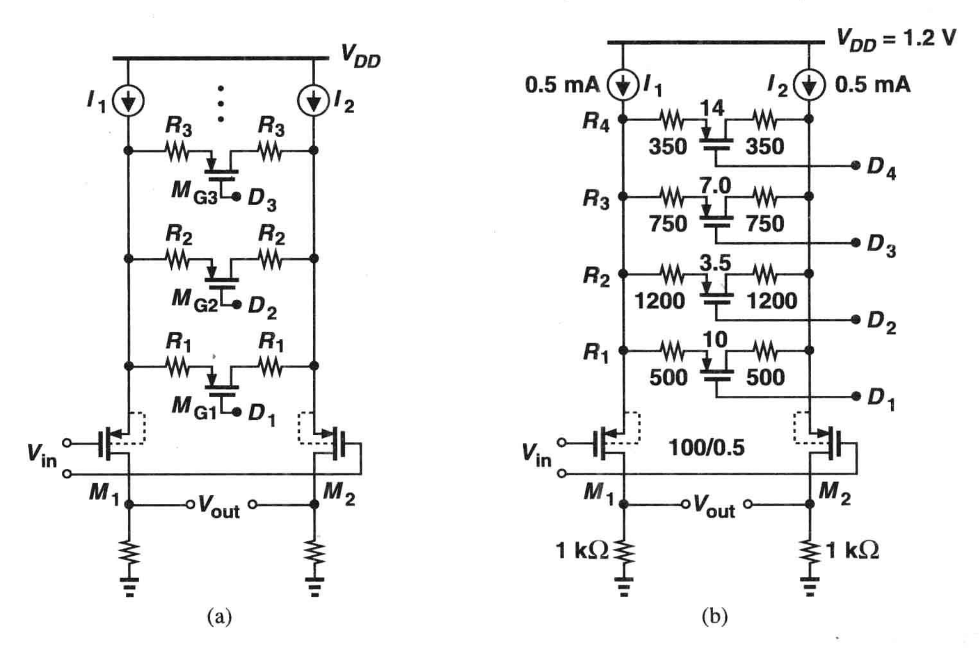


Figure 13.26 *(a) Simplified and (b) complete VGA circuit diagrams. (The n-wells are connected to V_{DD}.)*

In the circuit of Fig. 13.26(a), the nonlinearity of M_{G1}-M_{Gn} may manifest itself for large input swings. For this reason, these transistors must be wide enough that their on-resistance is only a fraction (e.g., one-tenth to one-fifth) of $2R_j$. The value of each R_j is chosen so as to provide a linear-in-dB gain characteristic.

Figure 13.26(b) shows the design in detail. For M_1 and M_2, we employ a long channel, reducing the nonlinearity due to their voltage-dependent output resistance, and tie their source and n-well, allowing a headroom of about 200 mV for I_1 and I_2 and minimizing their noise contribution (Problem 13.8). The degeneration branches provide a gain step of 2 dB.

Table 13.1 summarizes the simulated RX performance with the VGA placed after the chain. As with the topology of Fig. 13.22, the switches are driven by a "thermometer" code, i.e., $D_1D_2D_3D_4$ "fills" up by one more logical ONE for each 2-dB gain increase. We observe that (a) the RX P_{1dB} drops from -26 dBm to -31 dBm when the VGA is added to

13. Also called a "programmable-gain amplifier" (PGA).

Table 13.1 *Summary of receiver performance with gain switching.*

$D_1D_2D_3D_4$ (Fine AGC)	0000	0001	0011	0111
Gain (dB)	30	28	26	24
P_{1dB} (dBm)	–31	–30	–29	–28
NF (dB)	4.5	4.5	4.6	4.7

the chain, and (b) the noise figure rises by 0.2 dB in the low-gain mode. The VGA design thus favors the NF at the cost of P_{1dB}—while providing a maximum gain of 8 dB.

Example 13.17

A student seeking a higher P_{1dB} notes that the NF penalty for $D_1D_2D_3D_4 = 0011$ is negligible and decides to call this setting the "high-gain" mode. That is, the student simply omits the higher gain settings for 0000 and 0001. Explain the issue here.

Solution:

In the "high-gain" mode, the VGA provides a gain of only 4 dB. Consequently, the noise of the *next* stage (e.g., the baseband filter) may become significant.

13.3 TX DESIGN

The design of the TX begins with the power amplifier and proceeds backwards. The need for matching networks makes it extremely difficult to realize a PA operating in both 11g and 11a bands. We therefore assume two different PAs.

13.3.1 PA Design

As mentioned in Section 13.1.2, the PA must deliver +16 dBm (40 mW) with an output P_{1dB} of +24 dBm. The corresponding peak-to-peak voltage swings across a 50-Ω antenna are 4 V and 10 V, respectively. We assume an off-chip 1-to-2 balun and design a differential PA that provides a peak-to-peak swing of 2 V, albeit to a load resistance of 50 $\Omega/2^2 = 12.5\,\Omega$.[14] Figure 13.27 summarizes our thoughts, indicating that the *peak* voltage swing at X (or Y) need be only 0.5 V.

14. We neglect the loss of the balun here. In practice, about 0.5 to 1 dB of margin must be allowed for the balun's loss.

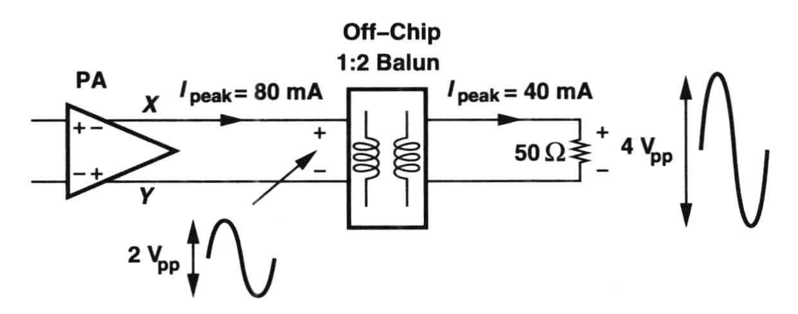


Figure 13.27 *Voltage swings provided by PA.*

Example 13.18

What P_{1dB} is necessary at node X (or Y) in Fig. 13.27?

Solution:

The balun lowers the P_{1dB} from +24 dBm (10 V_{pp}) across the antenna to 5 V_{pp} for V_{XY}. Thus, the P_{1dB} at X can be 2.5 V_{pp} (equivalent to +12 dBm).

The interesting (but troublesome) issue here is that the PA supply voltage must be high enough to support a single-ended P_{1dB} of 2.5 V_{pp} even though the actual swings rarely reach this level.

Let us begin with a quasi-differential cascode stage [Fig. 13.28(a)]. As explained in Chapter 12, the choice of V_b is governed by a trade-off between linearity and device stress. If V_b is too high, then the downward voltage swing at X and Y drives M_3 and M_4 into the triode region, causing the drain voltages of M_1 and M_2 to change and possibly create compression. If V_b is too low, then the upward voltage swing at X and Y produces an excessive drain-source voltage for M_3 and M_4.

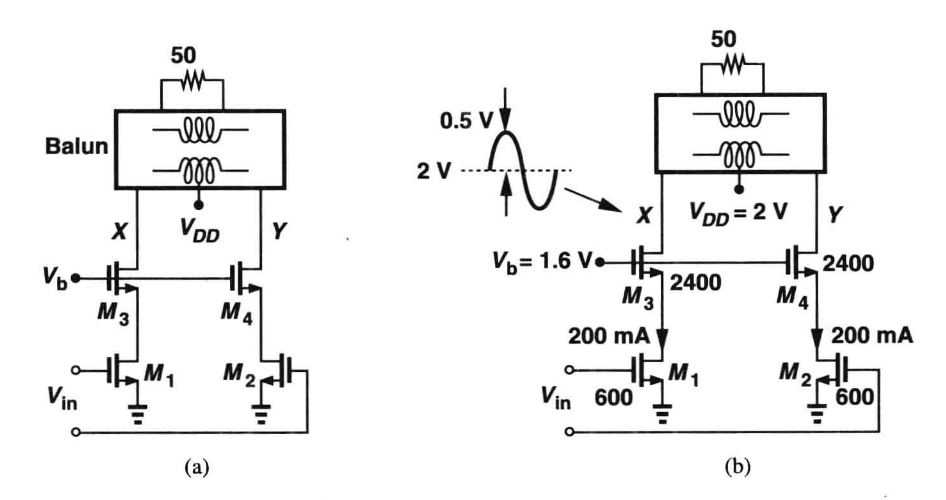


Figure 13.28 *(a) Simplified and (b) complete PA circuit diagrams.*

Another key principle in the design of the above stage is that the circuit must reach compression first at the *output* rather than at the input. To understand this point, suppose, for a given input swing, M_1 and M_2 experience compression in their I_D-V_{GS} characteristic while the output has not reached compression. (Recall that as the gate voltage of either transistor rises, its drain voltage falls, possibly driving the device into the triode region even if M_3 and M_4 are saturated.) This means that the supply voltage can be lowered without degrading the P_{1dB}. That is, if the input compresses first, then some of the voltage headroom chosen for the output is "wasted."

Another important principle is that the gain of the above stage must be maximized. This is because a higher gain translates to a lower input swing (for a given output P_{1dB}), ensuring that the circuit does not compress at the input first.

We also recognize that the single-ended load resistance seen at X (or Y) is equal to $50\ \Omega/2^2/2 = 6.25\ \Omega$. The circuit must therefore employ wide transistors and high bias currents to drive this load with a reasonable gain.

Example 13.19

Study the feasibility of the above design for a voltage gain of (a) 6 dB, or (b) 12 dB.

Solution:

With a voltage gain of 6 dB, as the circuit approaches P_{1dB}, the single-ended input peak-to-peak swing reaches $2.5\text{ V}/2 = 1.25\text{ V}$!! This value is much too large for the input transistors, leading to a high nonlinearity.

For a voltage gain of 12 dB, the necessary peak-to-peak input swing near P_{1dB} is equal to 0.613 V, a more reasonable value. Of course, the input transistors must now provide a transconductance of $g_m = 4/(6.25\ \Omega) = (1.56\ \Omega)^{-1}$, thus demanding a very large width and a high bias current.

Figure 13.28(b) shows the resulting design for a gain of 12 dB.[15] Fig. 13.29 plots the internal node voltage waveforms of the PA, and Fig. 13.30 depicts the compression characteristic and the (drain) efficiency as a function of the single-ended input level.

The design meets two criteria: (1) the gain falls by no more than 1 dB when the voltage swing at X (or Y) reaches 2.5 V_{pp}, (2) the transistors are not stressed for the average output swing, 1 V_{pp} at X (or Y). The cascode transistors are 2400 μm wide, reducing the voltage swing at the drains of M_1 and M_2. The input peak-to-peak voltage swing applied at the gate of M_1 (and M_2) is equal to 0.68 V when the output reaches P_{1dB}.

The above PA stage draws a total current of 400 mA from a 2-V supply, yielding an efficiency of about 30% at the output P_{1dB} and 5% at the average output level of 40 mW. This is the price paid for a back-off of 8 dB. More advanced designs achieve higher efficiencies [7, 8].

15. Note that the balun provides another 6 dB of voltage gain.

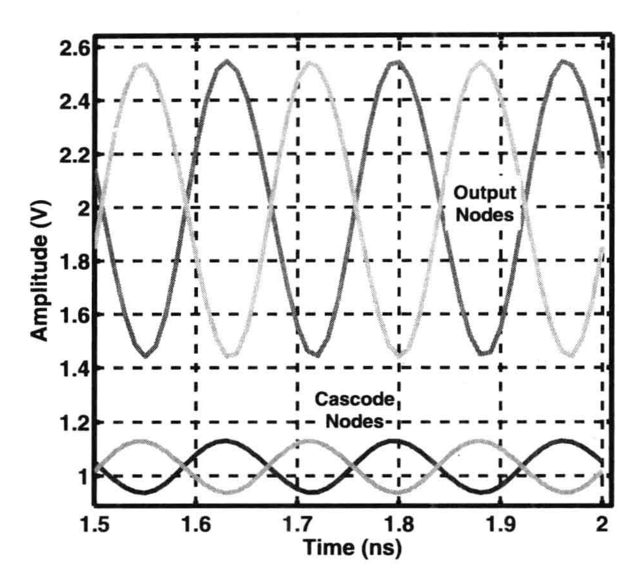


Figure 13.29 *PA waveforms.*

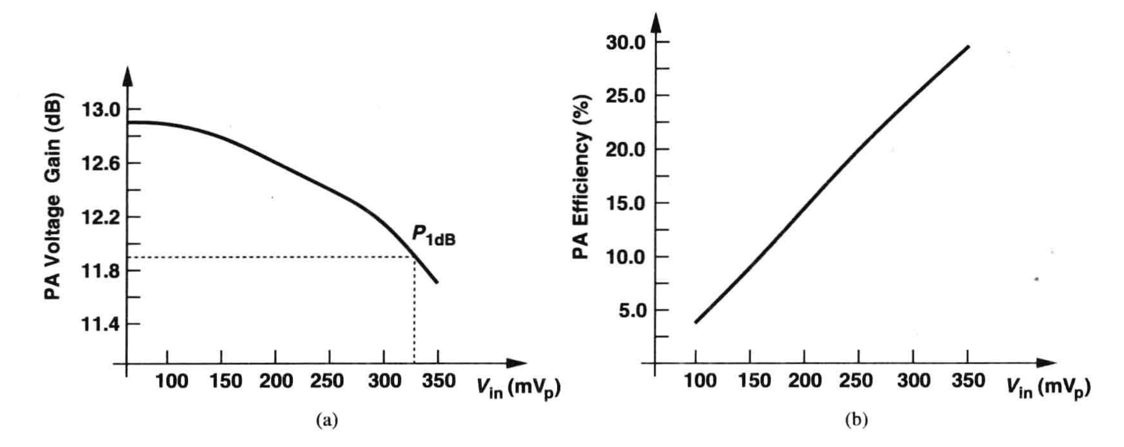


Figure 13.30 *PA's (a) compression characteristic and (b) efficiency.*

Predriver We now turn our attention to the PA predriver stage. The input capacitance of the PA is about 650 fF, requiring a driving inductance of about 1 nH for resonance at 6 GHz. With a Q of 8, such an inductor exhibits a parallel resistance of 300 Ω. The predriver must therefore have a bias current of at least 2.3 mA so as to generate a peak-to-peak voltage swing of 0.68 V. However, for the predriver not to degrade the TX linearity, its bias current must be quite higher.

Figure 13.31 shows the predriver and its interface with the PA. The width and bias current of M_5 and M_6 are chosen so as to provide a high linearity and a voltage gain of about 7 dB. The load inductor is reduced to 2×0.6 nH to accommodate the predriver parasitics. Resistor R_1 sustains a voltage drop of 0.5 V, biasing M_1 and M_2 at their nominal current. In practice, this resistor may be replaced with a tracking circuit to define this current more accurately.

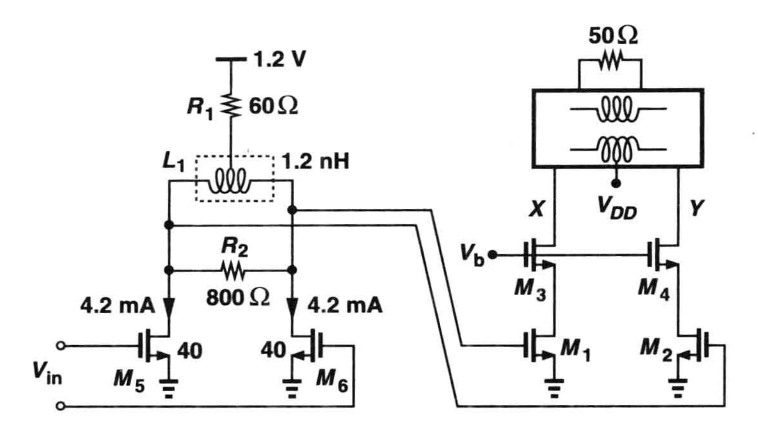


Figure 13.31 *PA predriver.*

Designed for resonance at 6 GHz with a Q of 8, the predriver suffers from a low gain at 5 GHz. Resistor R_2 is added to increase the bandwidth, but a few capacitors must be switched into the tank so as to lower the resonance frequency (Chapter 5). Inductor L_1 can also be raised so as to reduce the resonance frequency to about 5.5 GHz.

Example 13.20

A student decides to use ac coupling between the predriver and the PA so as to define the bias current of the output transistors by a current mirror. Explain the issues here.

Solution:

Figure 13.32 depicts such an arrangement. To minimize the attenuation of the signal, the value of C_c must be about 5 to 10 times the PA input capacitance, e.g., in the range of 3 to 6 pF. With a 5% parasitic capacitance to ground, C_p, this capacitor presents an additional load capacitance of 150 to 300 fF to the predriver, requiring a smaller driving inductance. More importantly, two coupling capacitors of this value occupy a large area.

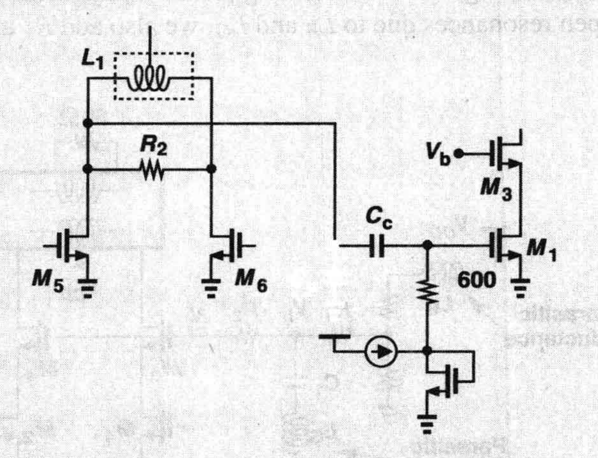


Figure 13.32 *Capacitive coupling between PA predriver and output stage.*

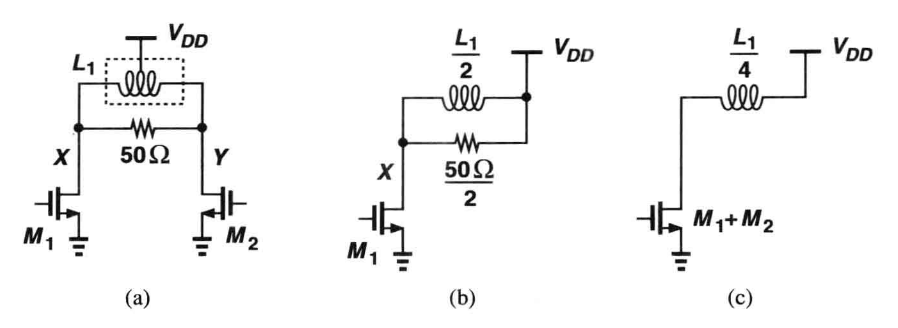


Figure 13.33 *(a) Stage driving a floating 50-Ω load, (b) half circuit for differential signals, and (c) half circuit for common-mode signals.*

Common-Mode Stability Quasi-differential PAs exhibit a higher common-mode gain than differential gain, possibly suffering from CM instability. To understand this point, let us first consider the simple stage shown in Fig. 13.33(a), where a quasi-differential pair drives a 50-Ω load. The circuit is generally stable from the standpoint of differential signals because, as evident from the half circuit in Fig. 13.33(b), the 25-Ω resistance seen by each transistor dominates the load, avoiding a negative resistance at the gate (Chapter 5).

For CM signals, on the other hand, the circuit of Fig. 13.33(a) collapses to that shown in Fig. 13.33(c). The 50-Ω resistor vanishes, leaving behind an inductively-loaded common-source stage, which can exhibit a negative input resistance. To ensure stability, a positive *common-mode* resistance must drive this stage.

Now consider the circuit of Fig. 13.31 again. For common-mode signals, resistor R_1 appears in series with the gate of $M_1 + M_2$, improving the stability. Of course, the cascode output stage also helps with the stability, minimizing the negative resistance seen at the gates of M_1 and M_2—but only if the gates of M_3 and M_4 are tied to a voltage source with a low impedance. In practice, however, this task proves difficult because of the parasitic inductance in series with V_{DD} or ground. We therefore provide the cascode gate bias through a lossy network as shown in Fig. 13.34. Here, we generate V_b by means of a simple resistive divider, but, to dampen resonances due to L_B and L_G, we also add R_1 and R_2. Note that the

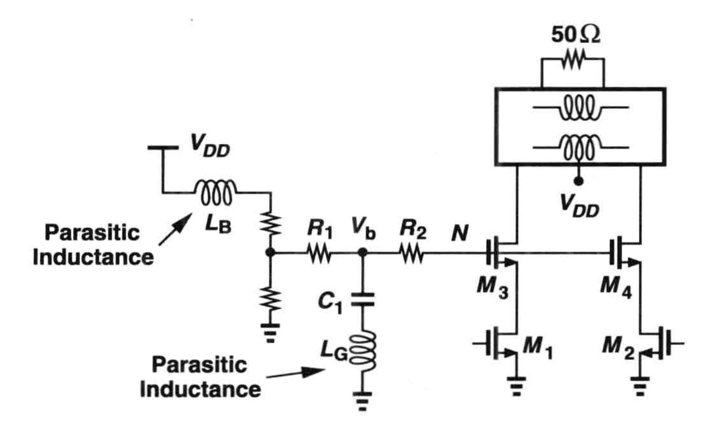


Figure 13.34 *Lossy network used to avoid CM instability.*

cascode operation remains intact for *differential* signals, i.e., node N appears as a virtual ground. This is another advantage of differential realizations.

13.3.2 Upconverter

The upconverter must translate the baseband I and Q signals to a 6-GHz center frequency while driving the 40-μm input transistors of the predriver. We employ a passive mixer topology here, assuming that rail-to-rail LO swings are available.

Figure 13.35 shows our first attempt at the upconverter construction and the necessary predriver modification. Each double-balanced mixer output voltage is converted to current, and the results are summed at nodes A and B. This arrangement must deal with two issues. First, since the gate bias voltage of M_5-M_8 is around 0.6 V, the mixer transistors suffer from a small overdrive voltage if the LO swing reaches only 1.2 V. We must therefore use ac coupling between the mixers and the predriver.

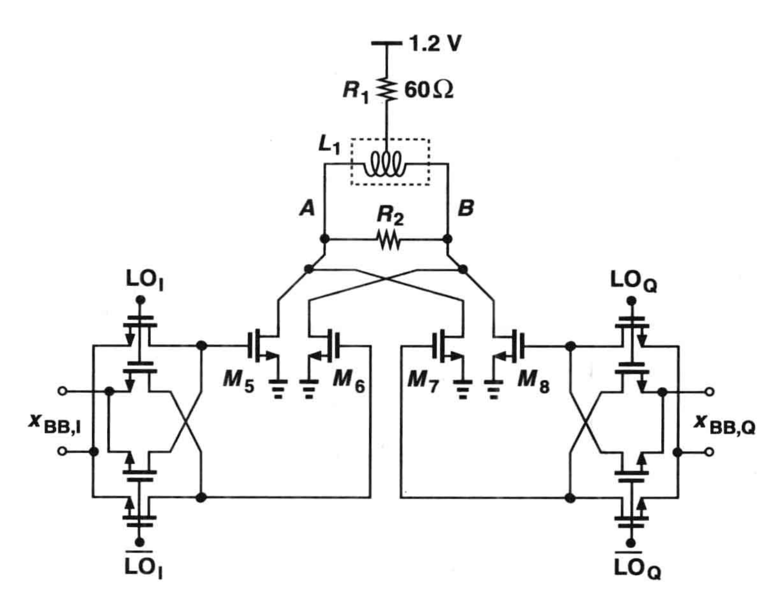


Figure 13.35 *Upconverter using passive mixers and V/I converters.*

Second, each passive mixer generates a *double-sideband* output, making it more difficult to achieve the output P_{1dB} required of the TX chain. To understand this point, consider the conceptual diagram in Fig. 13.36(a), where the TX is tested with a single baseband tone (rather than a modulated signal). The gate voltage of M_5 thus exhibits a beat behavior with a large swing, possibly driving M_5 into the triode region. Note that the drain voltage of M_5 has a constant envelope because the upconverted I and Q signals are summed at node A. The key point here is that, to generate a given swing at A, the beating swing at the gate of M_5 is *larger* than a constant-envelope swing that would be used to test only the predriver and the PA [Fig. 13.36(b)]. To overcome this difficulty, we wish to sum the signals *before* they reach the predriver.

Figure 13.37 shows the final TX design. Here, the mixer outputs are shorted to generate a single-sideband signal and avoid the beat behavior described above. This summation is possible owing to the finite on-resistance of the mixer switches. Simulations indicate

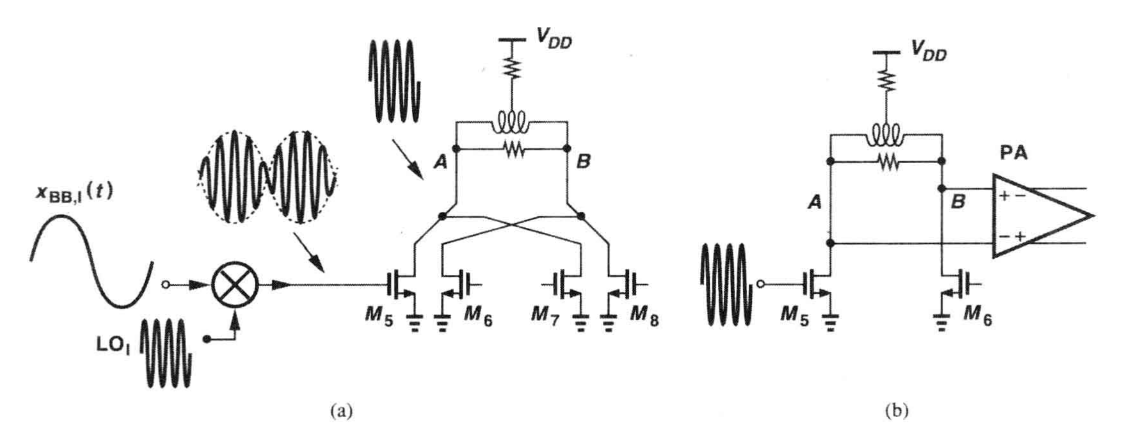


Figure 13.36 *(a) Problem of large beat swing at gate of V/I converter transistors, (b) stage without beat component.*

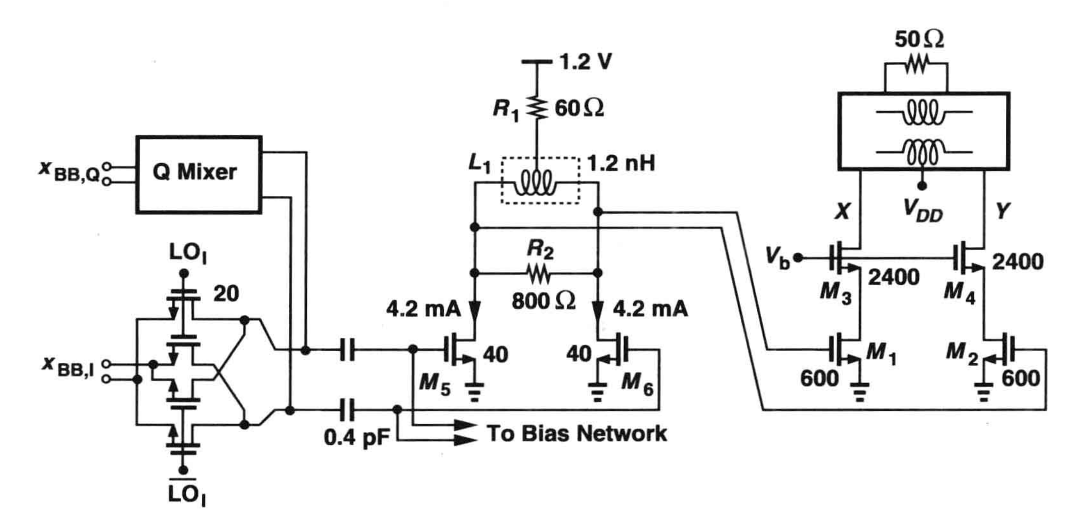


Figure 13.37 *Final TX design.*

that the gain and linearity of this upconverter topology are similar to those of the simple double-balanced counterpart. The baseband dc input of the mixers is around 0.3 V.

In order to determine the TX output P_{1dB}, we can plot the chain's conversion gain as a function of the baseband swing. The definition of the conversion gain is somewhat arbitrary; we define the gain as the differential voltage swing delivered to the 50-Ω load divided by the differential voltage swing of $x_{BB,I}(t)$ [or $x_{BB,Q}(t)$].

Figure 13.38 plots the overall TX conversion gain. The TX reaches its output P_{1dB} at $V_{BB,pp} = 890$ mV, at which point it delivers an output power of +24 dBm. The average output power of +16 dBm is obtained with $V_{BB,pp} \approx 350$ mV. The simulations assume a sinusoidal rail-to-rail LO waveform.

The large mixer transistors exhibit a threshold mismatch of 4 to 5 mV, resulting in some carrier feedthrough. A means of offset cancellation may be added to the stages preceding the mixers (usually I and Q low-pass filters) so as to suppress this effect.

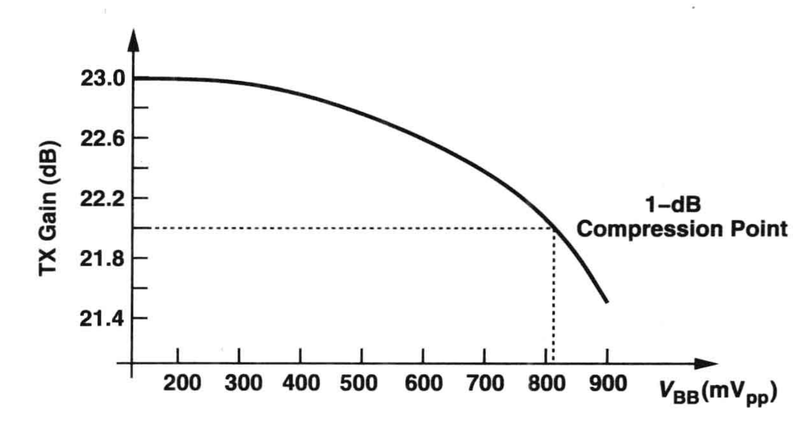


Figure 13.38 *TX compression characteristic.*

13.4 SYNTHESIZER DESIGN

In this section, we design an integer-N synthesizer with a reference frequency of 20 MHz for the 11a and 11g bands. From our analysis in Section 13.1.3, we must target an oscillator phase noise of about -101 dBc/Hz at 1-MHz offset for a carrier frequency of 2.4 GHz or 5 to 6 GHz. Recall from the frequency planning in Section 13.1.4 that the VCOs in fact operate at 10 to 12 GHz and must therefore exhibit a maximum phase noise of $-101 + 6 = -95$ dBc/Hz at 1-MHz offset.[16]

13.4.1 VCO Design

We choose the tuning range of the VCOs as follows. One VCO, VCO_1, operates from 9.6 GHz to 11 GHz, and the other, VCO_2, from 10.8 GHz to 12 GHz. The 200-MHz overlap between the VCOs' tuning ranges avoids a "blind zone" in the presence of modeling errors and random mismatches between the two circuits. We begin with VCO_2.

Let us assume a single-ended load inductance of 0.75 nH (i.e., a differential load inductance of 1.5 nH) with a Q of about 10 in the range of 10 to 12 GHz. Such values yield a single-ended parallel equivalent resistance of 618 Ω, requiring a tail current of about 1.5 mA to yield a single-ended peak-to-peak output swing of $(4/\pi)R_pI_{SS} = 1.2$ V. We choose a width of 10 μm for the cross-coupled transistors to ensure complete switching and assume a tentative load device width of 10 μm to account for the input capacitance of the subsequent frequency divider. Finally, we add enough constant capacitance to each side to obtain an oscillation frequency of about 12 GHz. Figure 13.39(a) shows this preliminary design.

At this point, we wish to briefly simulate the performance of the circuit before adding the tuning devices. Simulations suggest a single-ended peak-to-peak swing of about 1.2 V [Fig. 13.40(a)]. Also, the phase noise at 1-MHz offset is around -109 dBc/Hz [Fig. 13.40(b)], well below the required value. The design is thus far promising. However,

16. The phase noise of the frequency dividers following the VCOs is negligible.

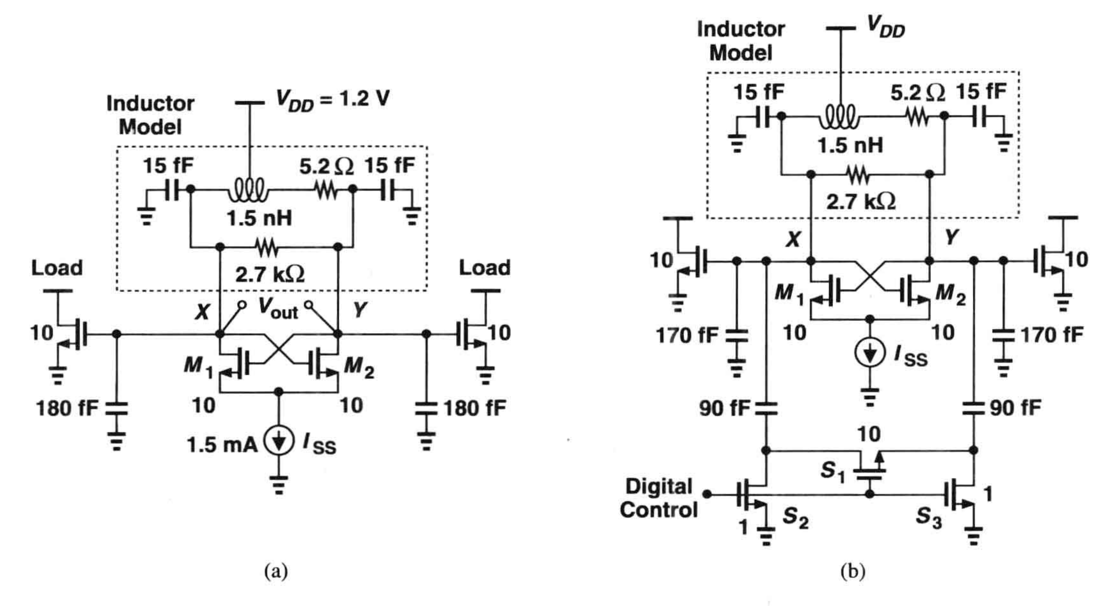


Figure 13.39 *(a) Preliminary 12-GHz VCO design, (b) addition of switched capacitors to lower the frequency to 10.8 GHz.*

as the drain and tail voltage waveforms suggest, the core transistors do enter the deep triode region, making the phase noise sensitive to the tail capacitance (Chapter 8).

Now, we add a switched capacitance of 90 fF to each side so as to discretely tune the frequency from 12 GHz to 10.8 GHz [Fig. 13.39(b)]. As explained in Chapter 8, the size of the switches in series with the 90-fF capacitors must be chosen according to the trade-off between their parasitic capacitance in the off state and their channel resistance in the on state. But a helpful observation in simulations is that the voltage swing decreases considerably if the on-resistance is not sufficiently small. That is, as the switches become wider, the swing is gradually restored.

Figure 13.39(b) depicts the modified design. We simulate the circuit again to ensure acceptable performance. Simulations indicate that the frequency can be tuned from 12.4 GHz to 10.8 GHz, but the single-ended swings fall to about 0.8 V at the lower end. As computed in Chapter 8, this effect arises from the sharp reduction of R_p (the parallel equivalent resistance of the tank) with frequency even if the switched capacitor branch does not degrade the Q. To remedy the situation, we raise the tail current to 2 mA. According to simulations, the phase noise at 1-MHz offset is now equal to -111 dBc/Hz at 10.8 GHz and -109 dBc/Hz at 12.4 GHz. We call S_1 a "floating" switch.

In the next step, we add varactors to the VCO and decompose the switched capacitors into smaller units, thus creating a set of discretely-spaced continuous tuning curves with some overlap. Note that the unit capacitors need not be equal. In fact, since at lower frequencies, the effect of a given capacitance change on the frequency is smaller (why?), we may begin with larger units at the lower end. This step of the design demands some iteration in the choice of the varactors' size and the number and values of the unit capacitors.

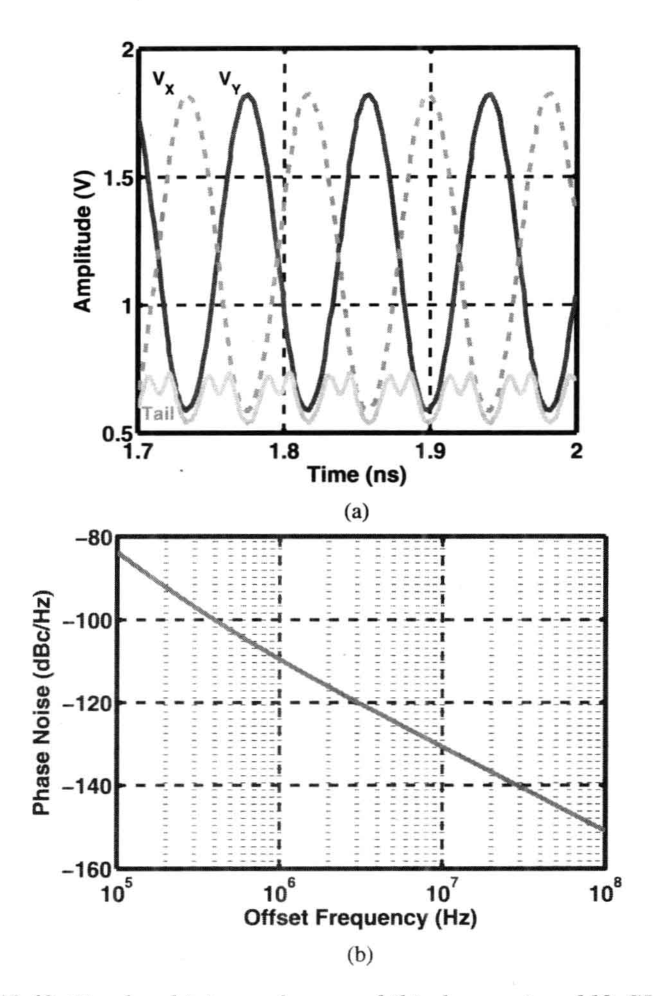


Figure 13.40 *Simulated (a) waveforms and (b) phase noise of 12-GHz VCO.*

After iterations, we arrive at the design in Fig. 13.41(a), where half of the circuit is shown for simplicity. Here, six switched capacitors and a 20-μm varactor provide the necessary tuning range. To reduce the frequency, first C_{u6} is switched in, then $C_{u6}+C_{u5}$, etc. We should make two remarks. First, as in Fig. 13.39(b), we still have floating switches even though they are not shown. Ideally, the switch widths are scaled with C_{uj}, but the minimum width in 65-nm technology is about 0.18 μm and is chosen for the grounded switches. The floating switches have a width of 2 μm for C_{u1} and C_{u2} and 1.5 μm for C_{u3}–C_{u6}.

Second, to obtain a wide continuous tuning range, the gate of the varactor is capacitively coupled to the core and biased at $V_b \approx 0.6$ V. As explained in Chapter 8, the "bottom-plate" parasitic of C_c may limit the tuning range. Fortunately, however, our design affords a large constant capacitance on each side, readily absorbing the parasitic of C_c. The coupling capacitor can be realized with parallel plates [Figure 13.41(b)]. The value of C_c is chosen about 10 times the maximum value of the varactor capacitance so as to negligibly reduce the tuning range.

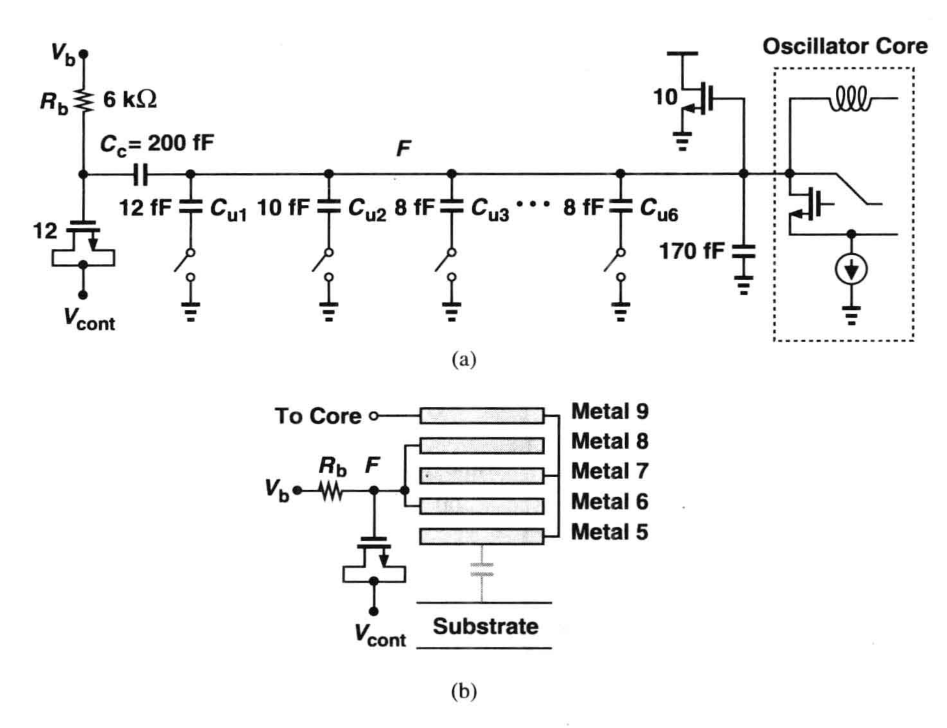


Figure 13.41 *(a) Switched-capacitor array added to VCO for discrete control and (b) coupling capacitor structure.*

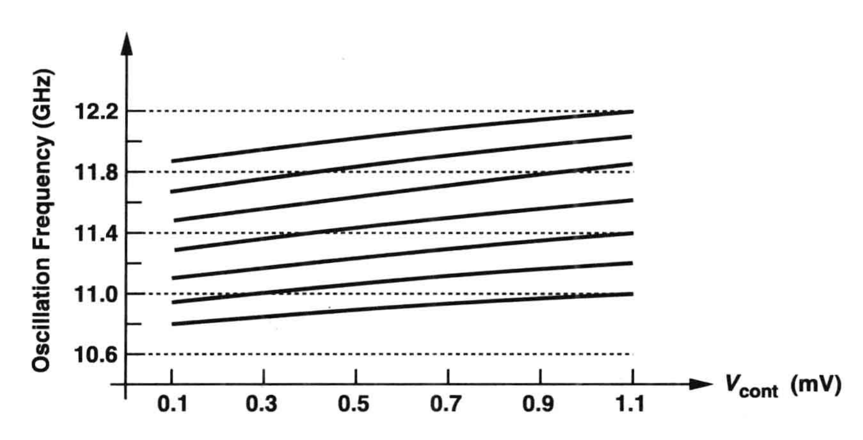


Figure 13.42 *VCO tuning characteristics.*

Figure 13.42 shows the VCO's tuning characteristics obtained from simulations. The control voltage is varied from 0.1 V to 1.1 V, with the assumption that the charge pump preceding the VCO can operate properly across this range. We note that K_{VCO} varies from about 200 MHz/V to 300 MHz/V. Figure 13.43 plots the phase noise with all of the capacitors switched into the tank.

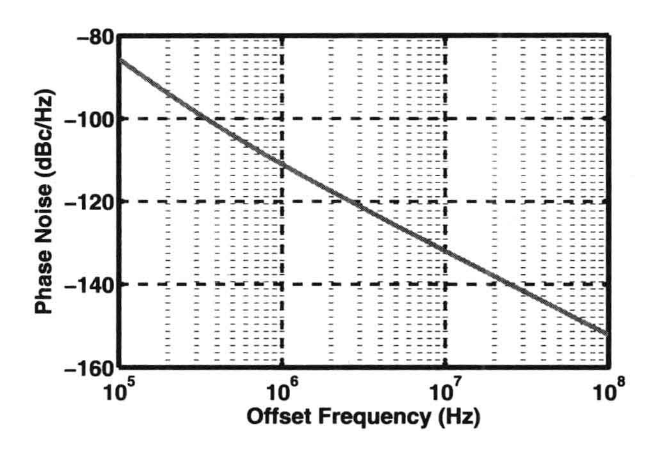


Figure 13.43 *VCO phase noise with all capacitors switched into the tank.*

Example 13.21

A student reasons that, since the thermal noise of R_b in Fig. 13.41(a) modulates the varactor voltage, the value of R_b must be *minimized*. Is this reasoning correct?

Solution:

Resistor R_b has two effects on the VCO: it lowers the Q of the tank and its noise modulates the frequency. We must quantify both effects.

Consider the simplified circuit shown in Fig. 13.44(a), where $L/2$, $R_p/2$, and C_T represent the single-ended equivalent of the tank (including the transistor capacitances and the switched capacitors). From Chapter 2, we know that C_c and C_{var} transform R_b to a value given by

$$R_{eq} \approx \left(1 + \frac{C_{var}}{C_c}\right)^2 R_b, \tag{13.34}$$

where the Q associated with this network is assumed greater than about 3. For $C_c \approx 10C_{var}$, we have $R_{eq} \approx 1.2R_b$. Thus, R_b must be roughly 10 times $R_p/2$ to negligibly reduce the tank Q.

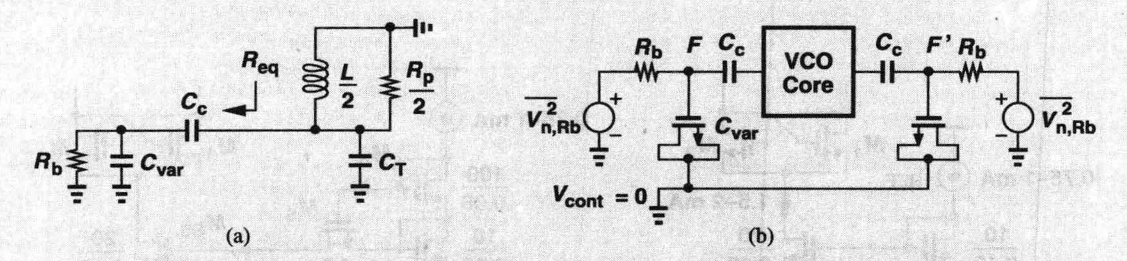


Figure 13.44 *(a) Equivalent tank impedance, (b) effect of noise of R_b.*

(Continues)

Example 13.21 *(Continued)*

But, can we use a *very* small R_b? In that case, the above expression for R_{eq} does not apply because the Q associated with C_c, C_{var}, and R_b is small. In the limit, as $R_b \to 0$, its effect on the tank Q vanishes again. However, the varactor is now "shorted out," failing to tune the frequency. We must therefore employ a large value for R_b.

Let us determine the phase noise due to R_b. The output phase noise of the VCO due to noise on the control voltage can be expressed as

$$S_{\phi n}(f) = S_{cont}(f)\frac{K_{VCO}^2}{\omega^2}, \tag{13.35}$$

where $S_{cont}(f)$ denotes the spectrum of the noise in V_{cont}. For offset frequencies below $\omega_{-3dB} \approx 1/(R_bC_c)$, the noise of R_b directly modulates the varactor, as if it were in series with V_{cont} [Fig. 13.44(b)]. To determine the phase noise with respect to the carrier, we make the following observations: (1) the gain from each resistor noise voltage to the output frequency is equal to $K_{VCO}/2$, where K_{VCO} denotes the gain from V_{cont} (Problem 13.9); (2) a two-sided thermal noise spectrum of $2kTR_b$ yields a phase noise spectrum around zero frequency given by $S_{\phi n} = 2kTR_b(K_{VCO}/2)^2/\omega^2$; (3) for an RF output of the form $A\cos(\omega_c t + \phi_n)$, the relative phase noise around the carrier is still given by $S_{\phi n}$; (4) the phase noise power must be doubled to account for the two R_b's. The output phase noise is thus equal to

$$S_{\phi n}(f) = \frac{kTR_bK_{VCO}^2}{4\pi^2 f^2}. \tag{13.36}$$

For $R_b = 6\ \text{k}\Omega$ and $K_{VCO} = 2\pi(300\,\text{MHz/V})$, $S_{\phi n}(f)$ reaches $-117\,\text{dBc/Hz}$ at 1-MHz offset, a value well below the actual phase noise of the VCO. If this contribution is objectionable, finer discrete tuning can be realized so as to reduce K_{VCO}.

In the last step of our VCO design, we replace the ideal tail current source with a current mirror. Shown in Fig. 13.45(a), this arrangement incorporates a channel length of 0.12 μm to improve the matching between the two transistors in the presence of a V_{DS} difference. The width of M_{SS} is chosen so as to create a small overdrive voltage, allowing the V_{GS} to

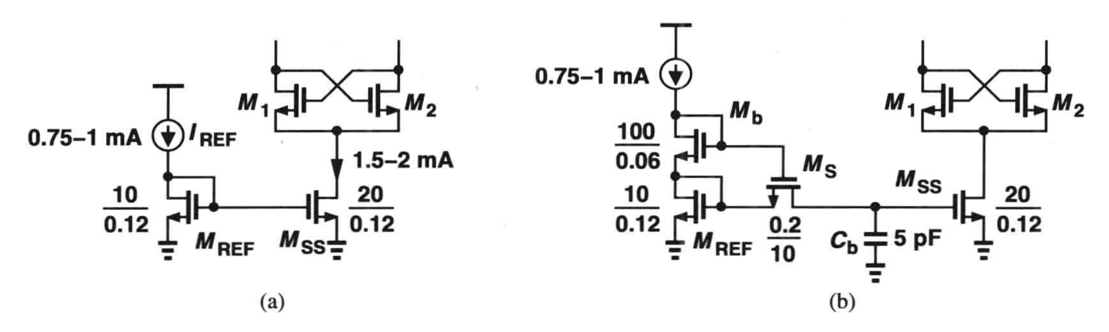


Figure 13.45 *(a) Current mirror used to bias the VCO, (b) modified mirror including a low-pass filter.*

be approximately equal to V_{DS} ($\approx$500 mV). This choice makes the transconductance and hence noise current of M_{SS} larger than necessary, but we will thus proceed for now. Note that M_{REF} and I_{REF} are scaled down by only a factor of 2 because the noise of M_{REF} may otherwise dominate.

The current mirror drastically raises the phase noise of the VCO, from -111 dBc/Hz to -100 dBc/Hz at 10.8 GHz and from -109 dBc/Hz to -98 dBc/Hz at 12.4 GHz (both at 1-MHz offset) (Fig. 13.46). According to Cadence, most of the phase noise now arises from the thermal and flicker noise of M_{REF} and M_{SS}.

A simple modification can suppress the contribution of M_{REF}. As shown in Fig. 13.45(b), we insert a low-pass filter between the two transistors, suppressing the noise of M_{REF} (and I_{REF}). To obtain a corner frequency well below 1 MHz, we (1) bias M_S with a small overdrive voltage, which is provided by the wide diode-connected transistor M_b; (2) select a width of 0.2 μm and a length of 10 μm for M_S; and (3) choose a value of 5 pF for C_b. The phase noise at 1-MHz offset is now equal to -104 dBc/Hz at 10.8 GHz and -101 dBc/Hz at 12.4 GHz. Figure 13.47 plots the phase noise of the final design. (The Q of the varactors is assumed to be high.)

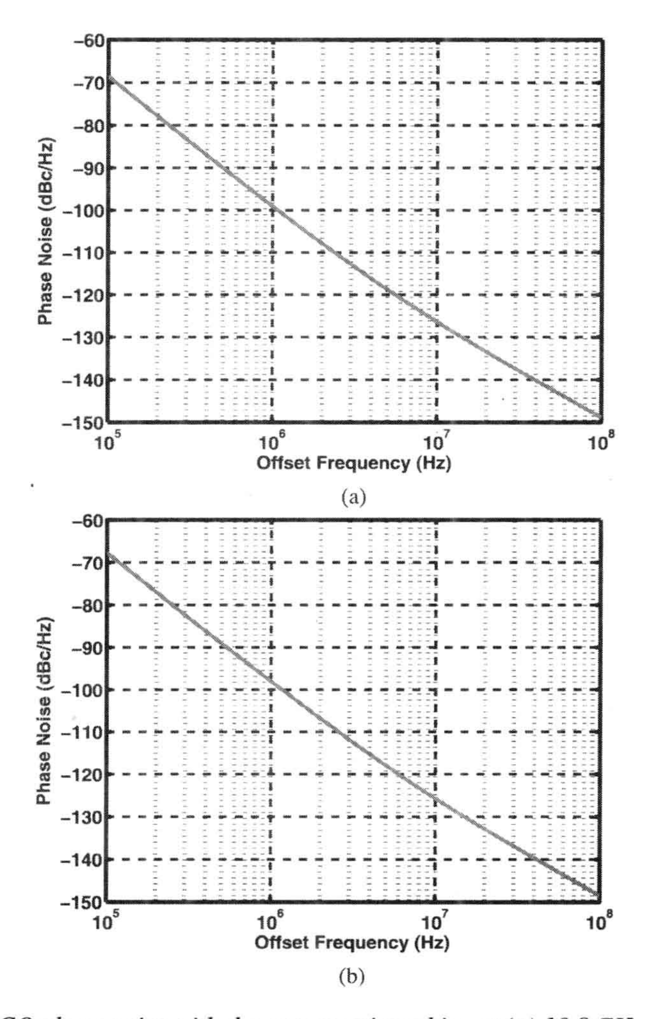


Figure 13.46 *VCO phase noise with the current mirror bias at (a) 10.8 GHz and (b) 12 GHz.*

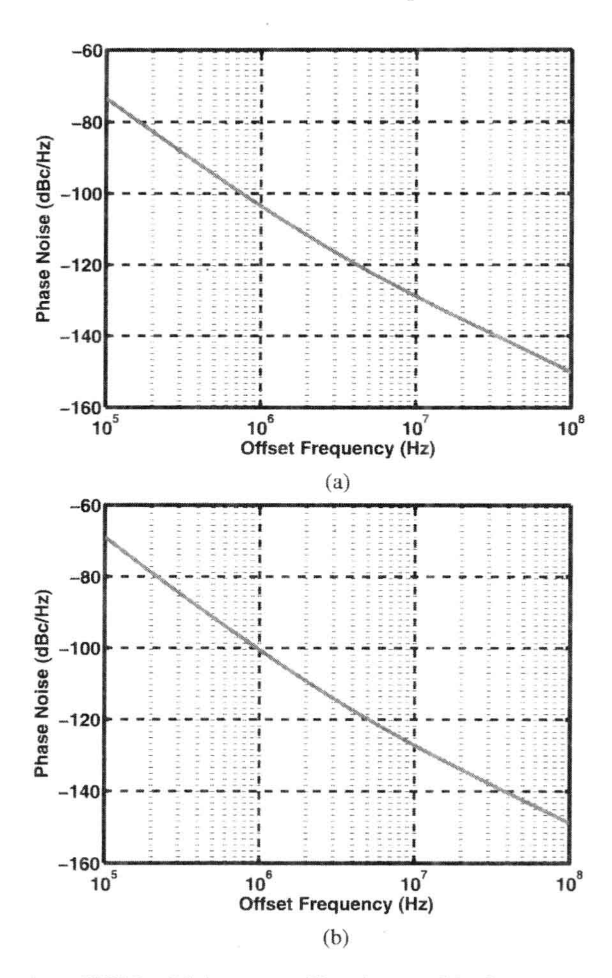


Figure 13.47 *Phase noise of VCO with low-pass filter inserted in the current mirror at (a) 10.8 GHz and (b) 12 GHz.*

Capacitor C_b in Fig. 13.45(b) occupies a relatively large area, even if realized as a MOSFET. One can make M_S longer and C_b smaller. Ultimately, however, the drain-source voltage drop of M_S due to the gate leakage current of M_{SS} becomes problematic.

While exceeding our phase noise target, the final VCO design still incurs significant noise penalty from the tail transistor, M_{SS}. The reader is encouraged to apply the tail noise suppression techniques described in Chapter 8.

The second VCO must cover a frequency range of 9.6 GHz to 11 GHz. This is readily accomplished by increasing the load inductor from 1.5 nH to 1.8 nH. The remainder of the design need not be modified.

Example 13.22

How does the synthesizer loop decide which VCO to use and how many capacitors to switch into the tank?

Example 13.22 *(Continued)*

Solution:

The synthesizer begins with, say, VCO_2 and all capacitors included in the tank. The control voltage, V_{cont}, is monitored by a simple analog comparator (Fig. 13.48). If V_{cont} exceeds, say, 1.1 V, and the loop does not lock, then the present setting cannot achieve the necessary frequency. One capacitor is then switched out of the loop and the loop is released again. This procedure is repeated (possibly switching to VCO_1 if VCO_2 runs out of steam) until lock is obtained for $V_{cont} \leq 1.1$ V.

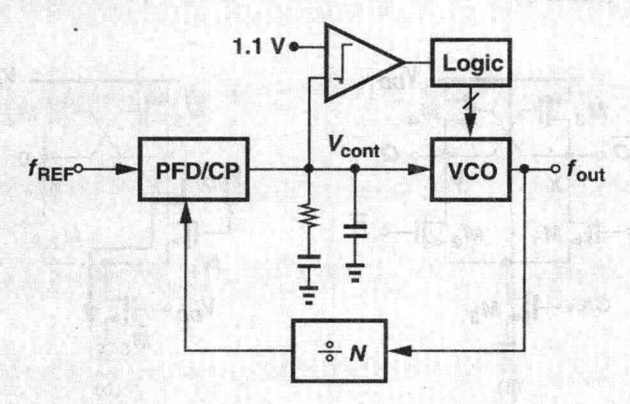


Figure 13.48 *Logic added to synthesizer for discrete tuning of VCO.*

The outputs of the two VCOs must be multiplexed. With rail-to-rail swings available, simple inverters can serve this purpose. As depicted in Fig. 13.49, each inverter is sized according to an estimated fanout necessary to drive the subsequent divide-by-2 circuit. The large transistors controlled by Select and $\overline{\text{Select}}$ enable one inverter and disable the other. Also, the feedback resistors bias the enabled inverter in its high-gain region. Note that the VCO outputs have a CM level equal to V_{DD} and are therefore capacitively coupled to the MUX.

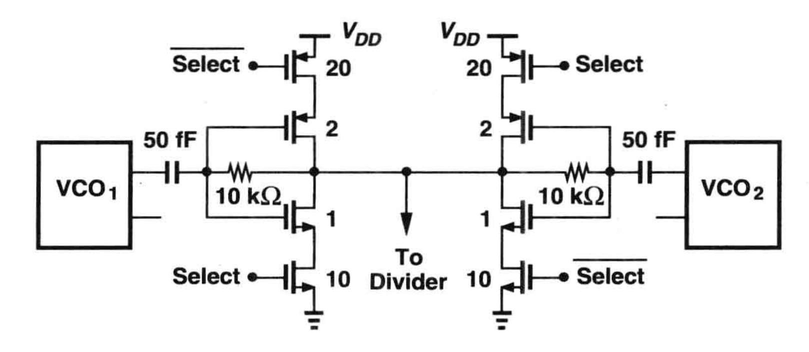


Figure 13.49 *Multiplexer selecting output of either VCO.*

13.4.2 Divider Design

Divide-by-2 Circuit The multiplexed VCO outputs must be divided by two so as to generate quadrature outputs. With rail-to-rail swings available at the MUX output, we seek a simple and efficient topology. The favorable speed-power trade-off of the Chang-Park-Kim divider described in Chapter 9 [9] makes it an attractive choice, but this topology does not produce quadrature (or even differential) phases.

Let us consider a complementary logic style that operates with rail-to-rail swings. Shown in Fig. 13.50(a) is a D latch based on this style. When CK is low, M_5 is off and the PMOS devices hold the logical state, when CK goes high, M_1 and M_2 force the input logical levels upon $\overline{\text{Q}}$ and Q.

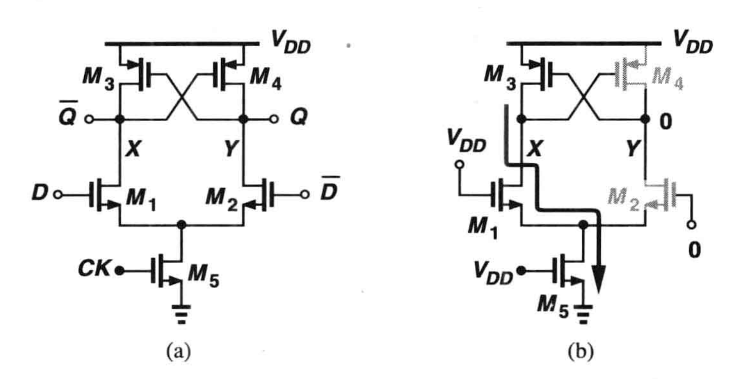


Figure 13.50 *(a) Latch topology, and (b) operation when one input goes high.*

The above circuit merits two remarks. First, this topology employs *dynamic* logic; as investigated in Problem 13.10, leakage currents eventually destroy the stored state if CK is low for a long time. Second, the latch is based on *ratioed* logic, requiring careful sizing. For example, if $\overline{\text{Q}}$ is high and CK goes high while $D \doteq 1$, then, as shown in Fig. 13.50(b), M_1 and M_5 appear in series and must "overcome" M_3. In other words, $R_{on1} + R_{on5}$ must be small enough to lower V_X to slightly below $V_{DD} - |V_{THP}|$ so that M_3 and M_4 can begin regeneration. In a typical design, $W_5 \approx W_{1,2} \approx 2W_{3,4}$. Speed requirements may encourage a wider M_5.

Example 13.23

The latch of Fig. 13.50(a) produces a low level *below* ground. Explain why.

Solution:

Suppose the clock has gone high and X and Y have reached ground and V_{DD}, respectively (Fig. 13.51). Now, the clock falls and is coupled through C_{GD5} to P, drawing a current from M_1 and hence X. Thus, V_X falls. If M_5 is a wide device to draw a large initial current, then this effect is more pronounced.

Example 13.23 *(Continued)*

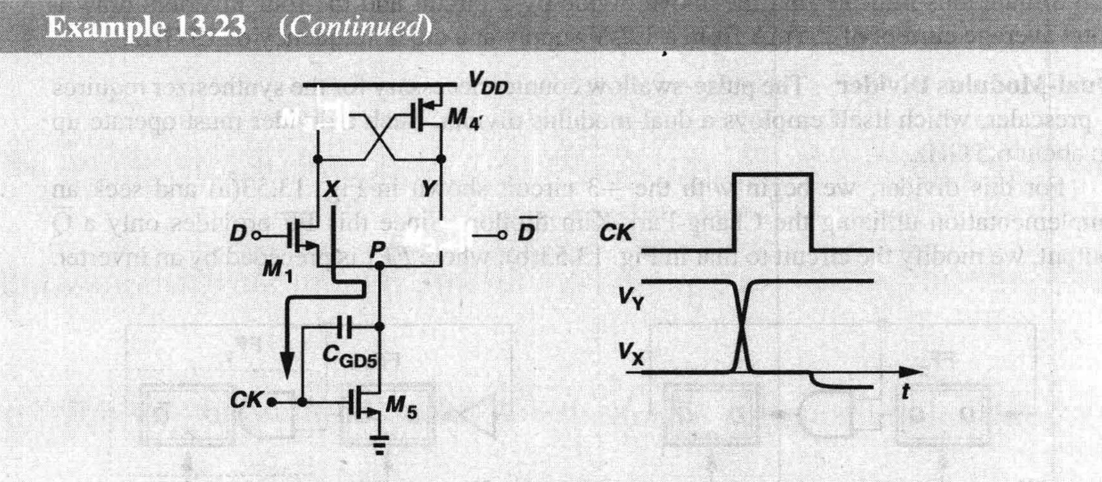


Figure 13.51 *Waveforms showing below-ground swings at the latch output.*

As with other latches, the above circuit may fail if loaded by a large load capacitance. For this reason, we immediately follow each latch in the divide-by-2 circuit by inverters. Figure 13.52 shows the result. The device widths are chosen for the worst case, namely, when the divider drives the TX passive mixers. The inverters present a small load to the latch but must drive a large capacitance themselves, thereby producing slow edges. However, the performance of the TX mixers is no worse than that predicted in Section 13.1.2, where the simulations assume a *sinusoidal* LO waveform.

Frequency dividers typically demand a conservative design, i.e., one operating well above the maximum frequency of interest. This is for two reasons: (1) the layout parasitics tend to lower the speed considerably, and (2) in the presence of process and temperature variations, the divider *must* handle the maximum frequency arriving from the VCO so as to ensure that the PLL operates correctly.

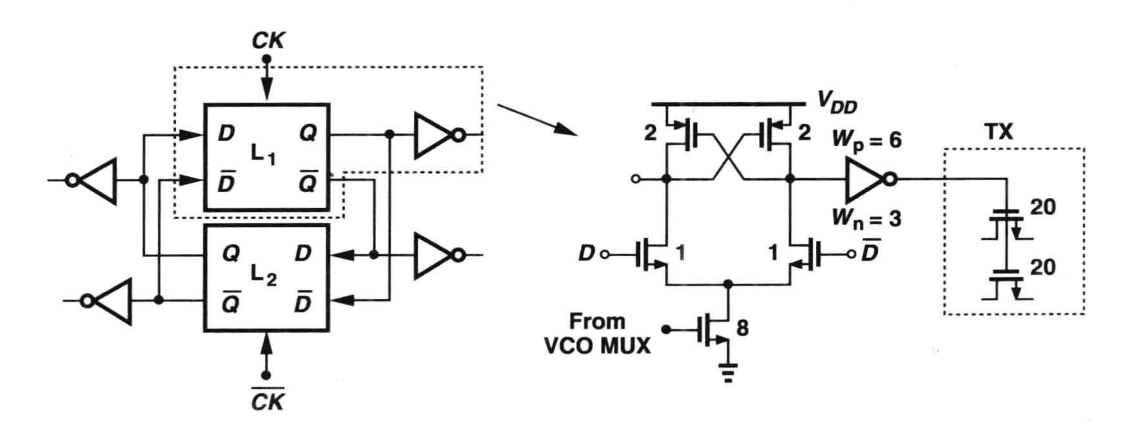


Figure 13.52 *Divide-by-two stage and its circuit details.*

Simulations indicate that the above divide-by-2 circuit and the four inverters draw a total average current of 2.5 mA from a 1.2-V supply at a clock frequency of 13 GHz.

Dual-Modulus Divider The pulse-swallow counter necessary for the synthesizer requires a prescaler, which itself employs a dual-modulus divider. Such a divider must operate up to about 6.5 GHz.

For this divider, we begin with the $\div 3$ circuit shown in Fig. 13.53(a) and seek an implementation utilizing the Chang-Park-Kim flipflop. Since this FF provides only a $\overline{Q}$ output, we modify the circuit to that in Fig. 13.53(b), where FF_1 is preceded by an inverter.

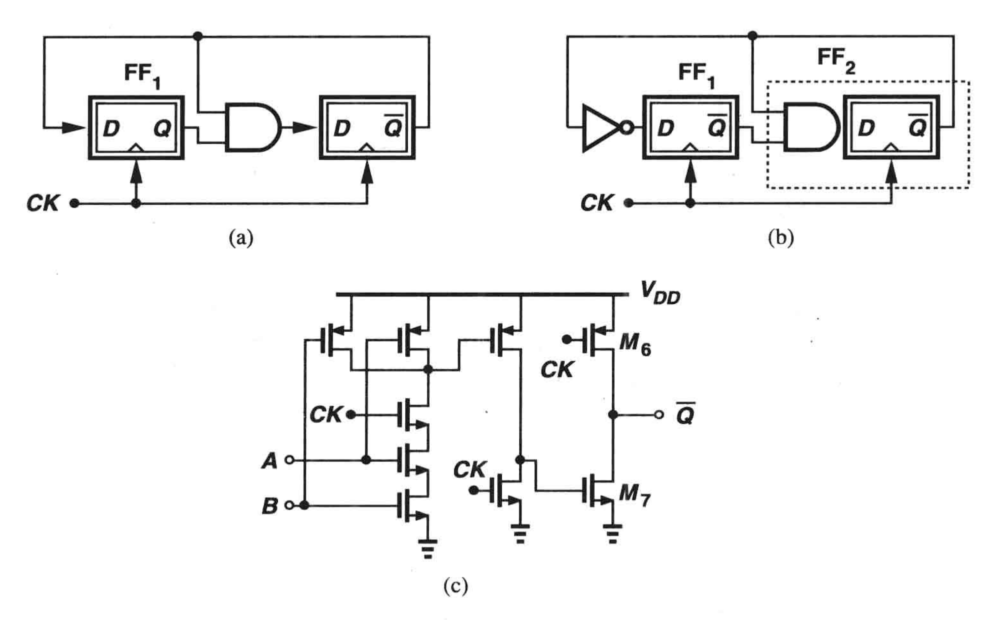


Figure 13.53 *(a) Dual-modulus divider with an explicit AND gate, (b) circuit of (a) with AND gate embedded within second flipflop, and (c) transistor-level implementation of AND and flipflop.*

We also wish to merge the AND gate with the second flipflop so as to improve the speed. Figure 13.53(c) depicts this AND/FF combination.

We must now add an OR gate to the topology of Fig. 13.53(a) to obtain a $\div 3/4$ circuit (Chapter 9). Again, we prefer to merge this gate with either of the flipflops. Figure 13.54 shows the overall $\div 3/4$ circuit design. The modulus control OR gate is embedded within the AND structure.

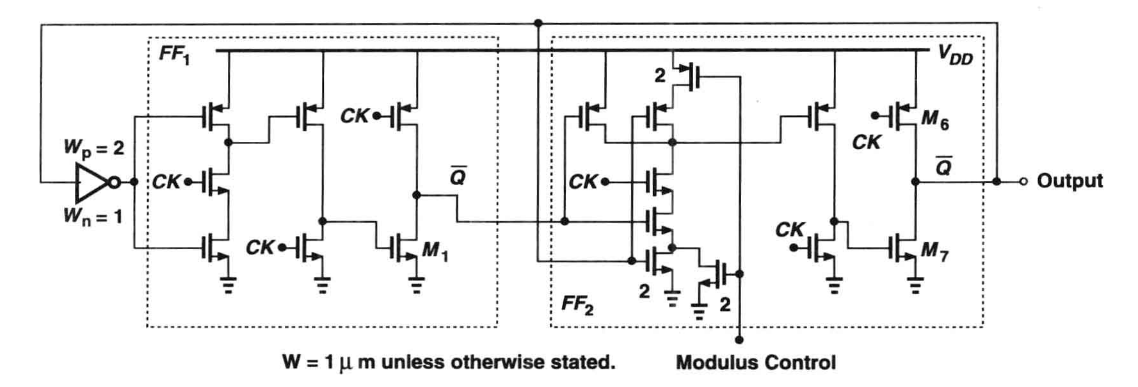


Figure 13.54 *Transistor-level implementation of dual-modulus prescaler.*

Plotted in Fig. 13.55 are the simulated output waveforms of the circuit in ÷4 and ÷3 modes at a clock frequency of 6.5 GHz. The divider draws 0.5 mA from a 1.2-V supply.

Example 13.24

A student observes that the circuit of Fig. 13.54 presents a total transistor width of 6 μm to the clock. The student then decides to halve the width of *all* of the transistors, thus halving both the clock input capacitance and the power consumption. Describe the pros and cons of this approach.

Solution:

This "linear" scaling indeed improves the performance. In fact, if the *load* seen by the main output could also be scaled proportionally, then the maximum operation speed would also remain unchanged (why?). In the present design, the 1-μm devices in the last stage of FF_2 drive $W = 4$ μm in the feedback path and can drive another 2 to 3 μm of load. A twofold scaling reduces the tolerable load to about 1 to 1.5 μm.

The ÷3/4 circuit can now be incorporated in a prescaler as described in Chapter 9. The reader is cautioned that the clock edge on which the asynchronous divide-by-2 stages change their outputs must be chosen carefully to avoid race conditions.

In order to cover a frequency range of 5180 to 5320 MHz in 20-MHz steps, the pulse-swallow counter must provide a divide ratio of $NP + S = 259$ to 266. If S varies from 9 to 16, then $NP = 250 = 5^3 \times 2$ and hence $N = 10$ and $P = 25$, requiring that the prescaler be designed as a ÷10/11 circuit. Alternatively, one can choose $N = 5$, $P = 50$, and a ÷5/6 prescaler.

The high 11a carrier frequencies, namely, from 5745 to 5805 MHz prove troublesome because they are not integer multiples of 20 MHz. An integer-N synthesizer must therefore operate with a reference frequency of 5 MHz, incurring a fourfold reduction in loop bandwidth. Our conservative VCO design in Section 13.4.1 still satisfies the free-running phase noise required of such a loop. The pulse-swallow counter must now provide $NP + S = 1149$ to 1161. For example, we can choose $S = 9$–21, $N = 10$, and $P = 114$, so that the above

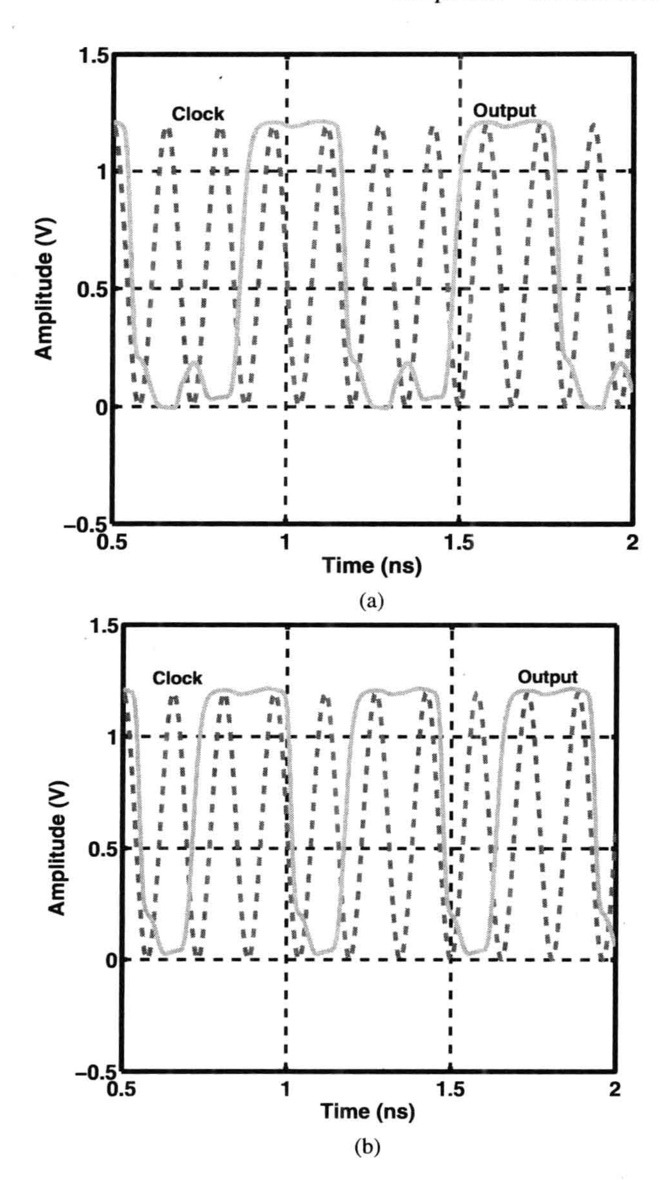


Figure 13.55 *Divider input and output waveforms for (a) divide-by-4 and (b) divide-by-3 operation.*

prescaler is utilized here as well. A fractional-N loop would be preferable here for accommodating the high band and other crystal frequencies with which the system may need to operate.[17] These designs are left as an exercise for the reader.

13.4.3 Loop Design

Let us now design the PFD/CP/LPF cascade and complete the synthesizer loop. The PFD is readily implemented using the NOR-based resettable latch topology described in Chapter 9.

17. For example, the crystal oscillator frequency may be dictated by the cell phone manufacturer or the baseband processor clock, etc.

The CP and LPF are designed based on the lowest value of K_{VCO} [$\approx 2\pi(200\,\text{MHz/V})$] and the highest value of the divide ratio, M (=2 × 1161 for a 5-MHz reference).

We begin with a loop bandwidth of 500 kHz and a charge pump current of 1 mA. Thus, $2.5\omega_n = 2\pi(500\text{ kHz})$ and hence $\omega_n = 2\pi(200\text{ kHz})$. We have

$$2\pi(200\text{ kHz}) = \sqrt{\frac{I_p K_{VCO}}{2\pi C_1 M}}, \tag{13.37}$$

obtaining $C_1 = 54.5$ pF. Such a capacitor occupies a large chip area. We instead choose $I_p = 2$ mA and $C_1 = 27$ pF, trading area for power consumption. Setting the damping factor to unity,

$$\zeta = \frac{R_1}{2}\sqrt{\frac{I_p K_{VCO} C_1}{2\pi M}} = 1, \tag{13.38}$$

yields $R_1 = 29.3$ kΩ. The second capacitor, C_2, is chosen equal to 5.4 pF.

For the charge pump, we return to the gate-switched topology described in Chapter 9 as it affords the maximum voltage headroom. Shown in Fig. 13.56, the design incorporates a channel length of 0.12 μm in the output transistors to lower channel-length modulation and wide devices tied to their gates to perform fast switching. To drive these devices, the PFD must be followed by large inverters.

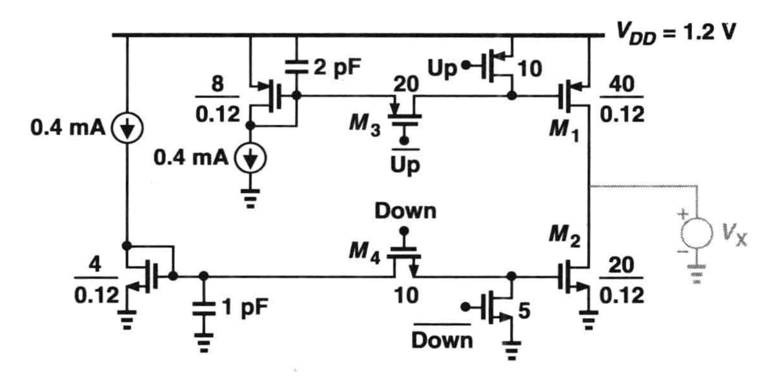


Figure 13.56 *Charge pump design.*

The gate-switched topology still proves rather slow, primarily because of the small overdrive of M_3 and M_4 in Fig. 13.56. That is, if the up and down pulses are narrow (so as to reduce the effect of mismatch between the up and down currents), then the gate voltages of M_1 and M_2 do not reach their final values, yielding output currents less than the target.

Figure 13.57 plots the simulated I/V characteristic of the charge pump. As explained in Chapter 9, in this test the Up and Down inputs are both asserted and a voltage source tied between the output node and ground is varied from V_{min} (=0.1 V) to V_{max} (1.1 V). Ideally equal to zero, the maximum current flowing through this voltage source reveals the deterministic mismatch between the Up and Down currents and the ripple resulting therefrom. In this design, the maximum mismatch occurs at $V_{out} = 1.1$ V and is equal to 60 μA, about 3%. If this mismatch creates an unacceptably large ripple, the CP techniques described in Chapters 9 and 10 can be employed.

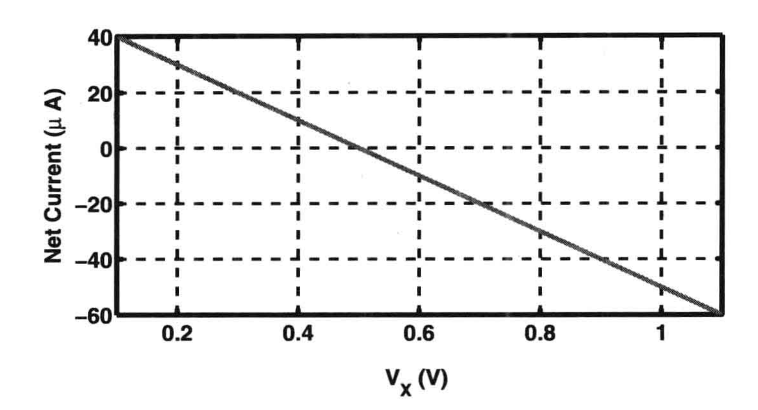


Figure 13.57 *Charge pump I/V characteristic.*

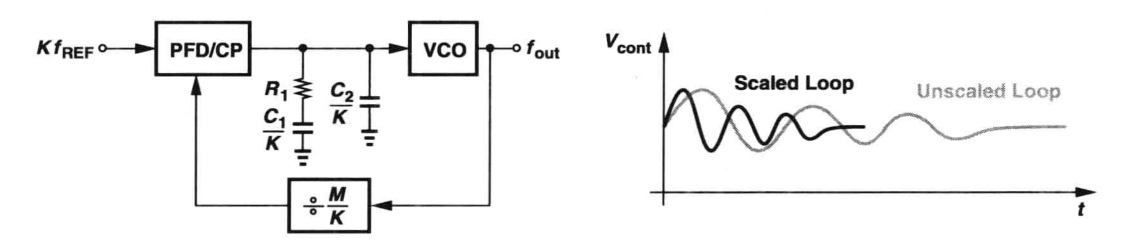


Figure 13.58 *Scaling loop parameters for time-contracted simulation.*

Loop Simulation The simulation of the synthesizer presents interesting challenges. With an input frequency of 5 MHz, the loop takes roughly 20 μs (100 input cycles) to lock. Moreover, for an output frequency of 12 GHz, the transient time step is chosen around 20 ps, requiring about one million time steps. Additionally, even without the discrete tuning logic of Fig. 13.48, the loop contains hundreds of transistors. Each simulation therefore takes several hours!

We begin the simulation by "time contraction" [6]. That is, we wish to scale down the lock time of the loop by a large factor, e.g., $K = 100$. To this end, we raise f_{REF} by a factor of K and reduce C_1, C_2, and M by a factor of K (Fig. 13.58). Of course, the PFD and charge pump must operate properly with a reference frequency of 500 MHz. Note that time contraction does not scale R_1, I_p, or K_{VCO}, and it retains the value of ζ while scaling down the loop "time constant," $(\zeta\omega_n)^{-1} = 4\pi M/(R_1 I_p K_{VCO})$, by a factor of K.

In addition to time contraction, we also employ a behavioral model for the VCO with the same value of K_{VCO} and f_{out}. The PFD, the CP, and the loop filter incorporate actual devices, thus producing a realistic ripple. Figure 13.59(a) shows the simulated settling behavior of the control voltage. The loop locks in about 150 ns, incurring a peak-to-peak ripple of nearly 30 mV [Fig. 13.59(b)]. We observe that our choice of the loop parameters has yielded a well-behaved lock response. This simulation takes about 40 seconds.

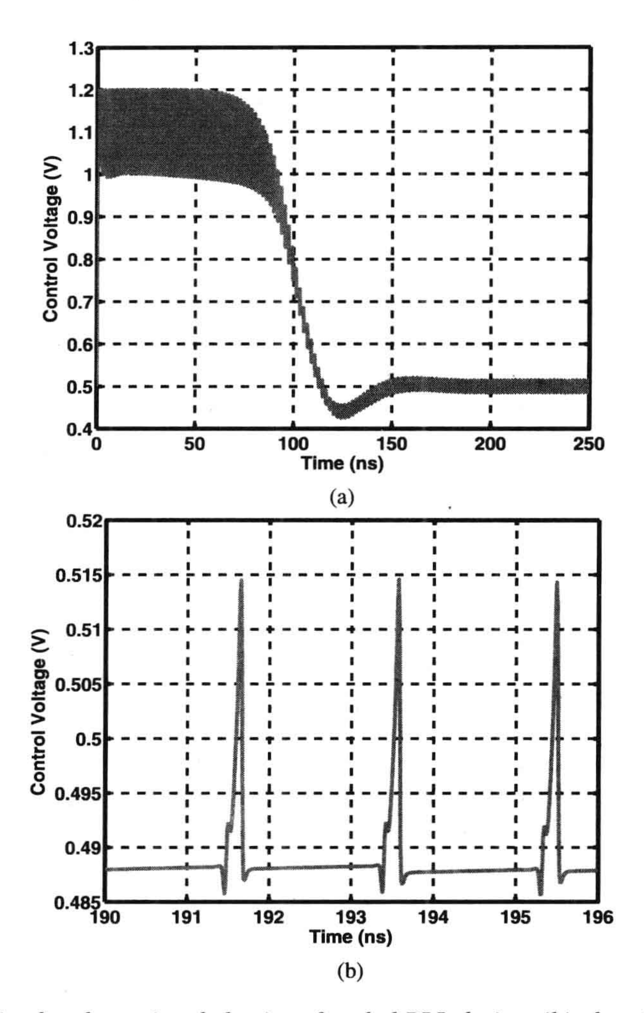


Figure 13.59 *(a) Simulated transient behavior of scaled PLL design, (b) plot of (a) for a narrower time scale showing the ripple waveform.*

Example 13.25

How is the control voltage ripple scaled with time contraction scaling?

Solution:

Since both C_1 and C_2 are scaled down by a factor of K while the PFD/CP design does not change, the ripple amplitude rises by a factor of K in the time-contracted loop.

The ripple revealed by the above simulation merits particular attention. Given that the amplitude falls 100-fold in the unscaled loop, we must determine whether the resulting sidebands at ±5-MHz offset have a sufficiently small magnitude. Recall from Chapter 9 that the ripple can be approximated by a train of impulses. In fact, if the area under

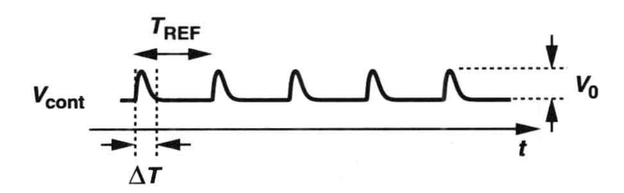


Figure 13.60 *Approximation of ripple by impulses.*

the ripple is given by, e.g., $V_0\Delta T$ (Fig. 13.60), then the relative magnitude of the sidebands is equal to $V_0\Delta TK_{VCO}/(2\pi)$. In the above simulations, the area under the ripple is roughly equal to $30\,\text{mV} \times 200\,\text{ps} \times 1/2$. This value is scaled down by a factor of 100 and multiplied by $K_{VCO}/(2\pi) = 200\,\text{MHz/V}$, yielding a relative sideband magnitude of $6 \times 10^{-4} = -64.4\,\text{dBc}$ at the output of the 12-GHz VCO. Thus, the 6-GHz carrier exhibits a sideband around $-70\,\text{dBc}$, an acceptable value.

REFERENCES

[1] L. L. Kan et al., "A 1-V 86-mW-RX 53-mW-TX Single-Chip CMOS Transceiver for WLAN IEEE 802.11a," *IEEE Journal of Solid-State Circuits,* vol. 42, pp. 1986–1998, Sept. 2007.

[2] K. Cai and P. Zhang, "The Effects of IP2 Impairment on an 802.11a OFDM Direct Conversion Radio System," *Microwave Journal,* vol. 47, pp. 22–35, Feb. 2004.

[3] I. Vassiliou et al., "A Single-Chip Digitally Calibrated 5.15-5.825-GHz 0.18-?m CMOS Transceiver for 802.11a Wireless LAN," *IEEE Journal of Solid-State Circuits,* vol. 38, pp. 2221–2231, Dec. 2003.

[4] C. Rapp, "Effects of HPA-Nonlinearity on a 4-DPSK/OFDM-Signal for a Digital Sound Broadband System," *Rec. Conf. ECSC,* pp. 179–184, Oct. 1991.

[5] M. Simon et al., "An 802.11a/b/g RF Transceiver in an SoC," *ISSCC Dig. Tech. Papers,* pp. 562–563, (also Slide Supplement), Feb. 2007.

[6] T.-C. Lee and B. Razavi, "A Stabilization Technique for Phase-Locked Frequency Synthesizers," *IEEE Journal of Solid-State Circuits,* vol. 38, pp. 888–894, June 2003.

[7] A. Afsahi and L. E. Larson, "An Integrated 33.5 dBm Linear 2.4 GHz Power Amplifier in 65 nm CMOS for WLAN Applications," *Proc. CICC,* pp. 611–614, Sept. 2010.

[8] A. Pham and C. G. Sodini, "A 5.8-GHz 47% Efficiency Linear Outphase Power Amplifier with Fully Integrated Power Combiner," *IEEE RFIC Symp. Dig. Tech. Papers,* pp. 160–163, June 2006.

[9] B. Chang, J. Park, and W. Kim, "A 1.2-GHz CMOS Dual-Modulus Prescaler Using New Dynamic D-Type Flip-Flops," *IEEE J. Solid-State Circuits,* vol. 31, pp. 749–754, May 1996.

PROBLEMS

13.1. Repeating the calculations leading to Eq. (13.7), determine the required IIP_3 of an 11a/g receiver for a data rate of 54 Mb/s and a sensitivity of $-65\,\text{dBm}$.

13.2. Suppose the interferers in Example 13.5 are not approximated by narrowband signals. Is the corruption due to reciprocal mixing greater or less than that calculated in the example?

13.3. Repeat Example 13.5 for the low sensitivity case, i.e., with the desired input at -65 dBm. Assume a noise-to-signal ratio of -35 dB.

13.4. Using the equations derived in Chapter 6 for the input impedance of a single-balanced voltage-driven passive mixer, estimate the load impedance seen by the LNA in Fig. 13.19.

13.5. Two blockers of equal power level appear in the adjacent and alternate adjacent channels of an 11a receiver. If the receiver has a phase noise of -100 dBc/Hz, what is the highest blocker level that allows a signal-to-noise ratio of 30 dB? Neglect other sources of noise.

13.6. Repeat the above problem for only one blocker in the adjacent channel and compare the results.

13.7. Assuming $\lambda > 0$, derive the voltage gain and input impedance of the LNA shown in Fig. 13.14(a).

13.8. Determine the noise contribution of I_1 and I_2 in Fig. 13.26(b) to the input for minimum and maximum gain settings. Neglect the on-resistance of the switches, channel-length modulation, and body effect.

13.9. In the circuit of Fig. 13.44(b), prove that the gain from the noise voltage of each resistor to the VCO output frequency is equal to K_{VCO}.

13.10. Considering the leakage current of the transistors in Fig. 13.50(a), prove that the state eventually vanishes if *CK* remains low indefinitely. Assuming each output node has a leakage current of I_1 and a total capacitance of C_1, estimate the time necessary for the state to vanish.

INDEX

B

C

D

H

I

J

L

O

P

S

T